WORLD KNOWLEDGE, CONTEXT AND STRUCTURE: THE USE AND ACQUISITION OF DIMENSIONAL ADJECTIVES

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

Sara Kessler
March 2019

© 2019 by Sara Nechama Kessler. All Rights Reserved.
Re-distributed by Stanford University under license with the author.



This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
http://creativecommons.org/licenses/by-nc/3.0/us/

This dissertation is online at: http://purl.stanford.edu/vg184qn9391

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

Meghan Sumner, Primary Adviser

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

Daniel Lassiter, Co-Adviser

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

Beth Levin

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

The interpretation of words in natural language is partially determined by the context in which they are used. Context can refer both to the immediate physical context and to the world knowledge of the speakers. It is known that both of these aspects of context play a role in language interpretation, but the contribution of each aspect, how the two aspects interact with each other and what constitutes world knowledge are open questions. The main goal of this dissertation is to examine these three questions. I focus on relative adjectives, such as *big* and *small*, whose meaning is typically evaluated relative to a contextually determined "comparison class". Their explicit context sensitivity makes these adjectives a good test case examining the role of context in language interpretation.

To accomplish this goal, I present a series of experiments aimed at showing the interaction of the immediate visual context and world knowledge in the interpretation of relative adjectives, as well as the interplay between different types of world knowledge in this process. The data from these experiments show that world knowledge is integrated with information from the immediate visual context and that at least two different types of world knowledge are involved. I argue that manipulating the stimuli present in the visual context affects the interpretation of the adjectives by shifting the salience of the two types and thus shifting the comparison class relative to which the adjectives are interpreted.

The secondary goal of this dissertation is to examine the role context plays in the acquisition of relative adjectives, focusing on a particular subset: dimensional adjectives, such as *big*, *tall* and *wide*. To address this question, I present an elicitation study of dimensional adjective antonyms. Dimensional adjectives all designate the

physical extent of an object; each specifies a particular dimension (e.g., verticality for *tall*) and polarity. Finally, each is associated with an antonym of opposing polarity; forming this association requires understanding the contribution of all of the adjectives' meaning components. Therefore, studies of antonym acquisition can provide a window into the structure of the semantic field. Through an experiment designed to elicit dimensional antonyms from Hebrew speaking children using pictures for contextual support, I show that context helps anchor children's association of one member of an antonym pair with the other. Their responses reflect the stages of acquisition of dimensional adjectives demonstrating progressive and structured knowledge of these meaning components. Together, the two investigations in this dissertation provide new insight from multiple perspectives on the interaction of context, world knowledge and language.

Acknowledgements

I am, above all, grateful to my committee, Meghan Sumner, Dan Lassiter and Beth Levin for their time and effort, discussions, insights and comments, reading of drafts and encouragement. They pushed me to make this dissertation the best it could be, and it is infinitely better for their feedback. Meghan agreed to take me on despite our work being in seemingly different fields, but we found common ground and her experimental rigor, excitement, and encouragement were instrumental in shaping this project. I am grateful to Dan for our many discussions of adjective semantics. His knowledge and critical perspective had a deep influence on my thinking, and his statistical help was much appreciated. I would like to thank Beth especially for her willingness to read and comment on draft after draft of pretty much anything I produced throughout my time at Stanford. I have learned so much from her about writing as well as linguistics and am grateful for all her support along the way. For guidance and advising, teaching and discussion, and inspiration for Chapter 5, I am grateful to Eve Clark.

I would also like to thank Tow Wasow and Judith Degen for sitting on my defense committee, asking insightful questions and providing helpful comments. I am also grateful to Judith for her help with the statistical analysis of the results in Chapters 2-4. Thanks also to Charlotte Fonrobert for being the external chair, as well as for her friendship and teaching during my time at Stanford.

My fascination with adjectives began in a classroom at Hebrew University in Jerusalem in 2012. Galit Sassoon's insightful and thought-provoking seminar, followed by the opportunity to edit her book, set me on this long and winding path. Without her mentorship, hours of enlightening discussion, attention to and love of

data, curiosity and friendship I would not be here today.

Various other linguistics faculty helped shape my academic thinking and interests. I would like to thank Chris Potts, John Rickford, Penny Eckert and Tom Wasow for their contributions, discussions, feedback, advice and questions. I would also like to thank my professors at Hebrew University for nurturing my interest in linguistics, and providing a solid theoretical foundation to build on. I am especially grateful to Edit Doron, Malka Rappaport-Hovav, Yael Ziv and Yehuda Falk z"l. Yehuda's true love of language and linguistics inspired a similar enthusiasm in his students, which I can only hope to aspire to.

I would like to thank the Stanford Linguistics Department for their financial support and acknowledge the Jewish Studies Department for funding a research assistantship in the summer of 2018.

For statistical help, javascript rescues, tech and TEXsupport, I am forever indebted to Simon Todd, Sebastian Schuster and my father, David Kessler. Thank you all for your time, patience and assistance at all hours, even from half a world away.

I am grateful to Elena Tribushinina for sharing her stimuli with me, enabling this project to get off the ground, and for work that inspires me to ask questions and dig for answers. I would also like to thank Mike Frank and the TAs of Psych 254, Cayce Hook and Eric Smith for feedback and help with setting up my original replication experiment and providing momentum for the rest of the dissertation. Additionally, I am grateful to all the anonymous MTurkers who did my experiments for their time, comments, and contribution to science.

My acquisition experiment would not have been possible without the children who participated and their parents who allowed it, took the time to bring their children to me, or welcomed me into their homes. I am particularly grateful to Hannah Aharon for her help in recruiting participants, and to Moria Ronen for hosting a massive playdate.

I cannot imagine my time at Stanford without my cohort, Daniel Galbraith, Olek Glowka, Sunwoo Jeong, Kate Lindsey, Prerna Nadathur and Simon Todd. We have been dubbed the Lunch Bunch for good reason, and these lunches, and the sunshine time they provided, were a highlight of my day. The camaraderie, discussions,

debates, warmth and support they provided, both social and academic, were so important to me, and are already missed. I am grateful to Simon, in particular, for the baking, the game nights, the conference buddying and the steadfast friendship. Beyond my cohort, I am grateful to my fellow grad students, especially James Collins, Tim Dozat, Roey Gafter, Lelia Glass, Sabrina Grimberg, Katherine Hilton, Masoud Jasbi, Ed King, Bonnie Krejci, Daisy Leigh, Dasha Popova and Sebastian Schuster for their friendship and collegiality, basement gossip sessions, discussions, trivia nights, coffee breaks, questions, insights and support.

Words cannot express how grateful I am to the Stanford Hillel community for providing me with a family and home away from home from my very first visit to Stanford for the department open house. That very first Shabbat at Stanford, the welcome I received and the community I saw, made it clear to me that this was a place I would be happy. I would like to thank the Hillel staff, past and present, including Shifra Elman, R. Jessica Kirschner, Lia Nilson, Jeremy Ragent, R. Daniel Silverstein and Elana Thurston-Milgrom for all the hard work they put into creating a space for community to flourish. I will always be grateful to Stanford Minyan for the opportunity to give back and for everything I got in return, and to Haim and Bracha Parke-Houben for always being there for all of us. Thanks to everybody who showed up and made it what it was, particularly Rachel, Zoe and Neil Barth, Arik, Gila, Daniel and Nili Bronstein, Michaela Elias, Michael Francus, David Glass, Josh and Goldie Grinberg, Dave Held, David Kahn, Irene Kaplow, Michal Leibowitz, Noah Linfield, Debra Linfield, David Lipton, Aaron Michelson, Annetta Parke-Houben, Yisroel Quint, Joe and Ariella Shayani, Dan Walker and Tamar Wenocur. Special shoutout to Cheli and Jack Ruth for taking over as chairpeople when I left. Thanks and yashar koach. I am especially grateful to all of the minyan babies over the years who let me play surrogate aunt and get the baby hugs I was missing from home (and to their parents who let me steal them). Last but not least, I couldn't have gotten through this whole process without Matan Leibovich and Raviv Murciano-Goroff cheering me on, keeping me company, work partying, coffee-breaking, dish walking, Sunday night dinnering, krispy-kremeing and generally being awesome people.

The greater Palo Alto Jewish community was also incredibly welcoming and generous. I would especially like to mention Noam and Maya, Jon and Beruria, Ari and Liora, Shani and Chaim, Jonah and Ziva and Ilana and David and their families for all of the shabbos meals, Purim seudahs, kickball games, leftovers, and learning. Special thanks to Ari Tuchman for unbeatable Sunday night shiurs. Michael and Ronna Greenwald and Bach treated me like family and gave me a second home in Berkeley, whenever I needed to get away. I couldn't have asked for a better aunt and uncle in-law.

While I was on the other side of the planet, I always had a support network back home providing encouragement, news, postcards and baby pictures. My friends from way back, who were always there over Skype, Whatsapp, Gchat, or when I made it back for visits: Elisheva Gordon, Adi Isaacson, Orly Jackel, Katie Michaels, Anna Miller, Avital Moshe, Michal Novoselsky, Avner and Yonit, Maayan and Tamir Segal. And finally, I would like to thank my family. They have made me who I am, in every sense of the word, and I couldn't be more grateful. My parents Judy and David Kessler, have been there for me, always. My siblings and their spouses Benjy and Michal Kessler and Naomi and Tomer Shamul, always get me and my niece and nephews, Adi, Nadav and Yonatan are the best baby faces around. Watching them grow from a distance was incredibly hard, and being back to watch it happen is a privilege and a pleasure. They make me smile by being there and have kept me sane throughout these last few months of writing. תושלב"ע

Contents

Abstract **iv**

Acknowledgements **vi**

1 Relative Adjective Semantics **1**

1.1 Introduction 1

1.2 Relative Adjective Semantics 4

1.3 Theoretical Framework for Comparison Classes 8

1.4 Previous Experimental Work 11

1.4.1 Experiments on Standard Setting 12

1.4.2 Comparison Class Experiments 19

1.5 Categories, World Knowledge and Comparison Classes 24

1.6 Roadmap 27

2 The Size of Things: A norming study **29**

2.1 Introduction 29

2.2 The Stimuli 30

2.2.1 The Stimuli for Experiment 1a 31

2.2.2 The Stimuli for Experiment 1b 32

2.2.3 The Stimuli for Experiments 2 and 3 32

2.2.4 The Stimuli for Experiments 4 and 5 32

2.3 Method 33

2.3.1 Participants 33

2.3.2 Materials and Procedure 33

2.3.3 Exclusions . . . 34
2.4 Results . . . 38
2.5 Discussion . . . 42

3 How World Knowledge Shifts Adjective Interpretation 44
3.1 Introduction . . . 44
3.2 Experiment 1a: Replication and method validation . . . 50
3.2.1 Methods . . . 51
3.2.2 Results . . . 55
3.2.3 Discussion . . . 59
3.3 Experiment 1b: Bunnies and ducklings . . . 61
3.3.1 Methods . . . 62
3.3.2 Results . . . 63
3.3.3 Discussion . . . 66
3.4 Experiment 2: Photographs . . . 68
3.4.1 Methods . . . 71
3.4.2 Results . . . 72
3.4.3 Discussion . . . 76
3.5 Experiment 3: Randomized order . . . 76
3.5.1 Methods . . . 77
3.5.2 Results . . . 78
3.5.3 Discussion . . . 83
3.6 General Discussion . . . 85

4 World Knowledge or Scale Structure Asymmetry? 90
4.1 Introduction . . . 90
4.2 Experiment 4: Tiny objects . . . 92
4.2.1 Methods . . . 93
4.2.2 Results . . . 94
4.2.3 Discussion . . . 97
4.3 Experiment 5: Randomized tiny objects . . . 98
4.3.1 Methods . . . 98

4.3.2 Results . . . 100
4.3.3 Discussion . . . 103
4.4 General Discussion . . . 103

5 The Structured Acquisition of Dimensional Adjective Antonyms 105
5.1 Introduction . . . 105
5.2 Background . . . 106
5.2.1 Dimensional Adjective Meaning . . . 106
5.2.2 Antonymy . . . 108
5.2.3 Hebrew Dimensional Adjectives . . . 109
5.2.4 The Acquisition of Dimensional Adjectives . . . 110
5.3 Antonym Acquisition Experiment . . . 116
5.3.1 Methods . . . 116
5.3.2 Results . . . 120
5.3.3 Discussion . . . 133
5.4 Conclusion . . . 136

6 Discussion 138
6.1 Introduction . . . 138
6.2 Summary . . . 139
6.3 Theoretical Implications . . . 145
6.3.1 Generalizability across Objects and Images . . . 146
6.3.2 Generalizability across Adjectives . . . 148
6.3.3 Generalizability to Other Domains . . . 152
6.4 Methodological Implications . . . 156
6.4.1 Implications for Visual World Experiments . . . 157
6.4.2 Implications for Standard of Comparison Experiments . . . 158
6.5 Conclusion . . . 160

List of Tables

2.1 Stimulus objects grouped by prototypicality status 31
2.2 Stimuli for each prototypicality status 35
2.3 Results for model with interaction between prototypicality status and stimulus type 40

3.1 Prototypicality status of the objects displayed in Experiment 1a . . . 53

5.1 Examples of the meaning compononents of dimensional adjectives . . 107
5.2 The Hebrew adjective paradigm for *gadol* ('big') 110
5.3 The adjectives tested and the nouns depicted for each item 118
5.4 Distribution of response types 121
5.5 Mean scores by age group for each adjective pair 124
5.6 Substitutions occuring more than once for positive polarity adjectives 127
5.7 Substitutions occuring more than once for negative polarity adjectives 127

6.1 Comparison of the mean *big* and *small* zones in Experiments 1–5 . . . 143

List of Figures

1.1 Sample stimulus set from Solt and Gotzner (2012, p. 3, Fig. 1) 13
1.2 Stimuli from Hansen and Chemla (2017, p. 252, Figs. 5 and 6) 15
1.3 Sample stimulus from Tribushinina (2011, p. 539, Fig. 1) 22
1.4 The mean number of images labeled *groot* ('big') and *klein* ('small') across prototypicality statuses, from Tribushinina (2011, p. 541, Fig. 2) 23

2.1 Sample trial display in the clipart condition, where the participant is asked to rate the size of a clipart mouse 34
2.2 Violin plot of ratings by prototypicality status category in each condition 35
2.3 Density curves of ratings by participant for deviating participants . . 36
2.4 Density curves of ratings by participant for eight arbitrarily chosen participants who had fewer than six deviant trials 37
2.5 Mean ratings of the objects by prototypicality status in each condition 38
2.6 Mean ratings of the objects by noun in each condition 39
2.7 Mean ratings of bunny, mouse, chick and duckling by noun in each condition 41
2.8 Mean ratings of the objects by prototypicality status in each condition for target objects, excluding mice and chicks 42

3.1 Sample trial display for Experiment 1a, showing hippos in ascending order 52
3.2 Mean number of images selected as *big* and *small* for each verb in Experiment 1a 56

3.3 Mean number of images selected as *big* and *small* by prototypicality status of the object in Experiment 1a 57
3.4 Mean number of images selected as *big* (left) and *small* (right) by object in Experiment 1a 59
3.5 The bunny and duckling images used in Experiment 1b in place of the mouse and chick images 62
3.6 Mean number of images selected as *big* and *small* by prototypicality status of the object in Experiment 1b 64
3.7 Mean number of images selected as *big* (left) and *small* (right) by object in Experiment 1b 65
3.8 Two sample stimulus sets; Display A shows clipart stimuli depicting elephants while Display B shows photographic stimuli of elephants. . 66
3.9 Sample trial display for Experiment 2, showing hippos in descending order 71
3.10 Mean number of images selected as *big* and *small* by prototypicality status of the object in Experiment 2 73
3.11 Mean number of images selected as *big* (left) and *small* (right) by object in Experiment 2 74
3.12 Number of images selected by each participant in each trial throughout Experiment 2 75
3.13 Sample trial display for Experiment 3, showing hippos in randomized order 78
3.14 Mean number of images selected as *big* and *small* by prototypicality status of the object in Experiment 3 79
3.15 Mean number of images selected as *big* (left) and *small* (right) by object in Experiment 3 81
3.16 Number of images selected by each participant in each trial throughout Experiment 3 82

4.1 Sample trial display for Experiment 4, showing thumb tacks in descending order 94

4.2 The mean *big* (red) and *small* (blue) zones by object in Experiment 4 95
4.3 Number of images selected by each participant in each trial throughout Experiment 4 96
4.4 Sample trial display for Experiment 5, showing thumb tacks in randomized order 99
4.5 The mean *big* (red) and *small* (blue) zones by object in Experiment 5 100
4.6 Number of images selected by each participant in each trial throughout Experiment 5 101

5.1 A sample stimulus scene for the adjective pair *big/small*, where the target for elicitation was *small*. 119
5.2 The percent of SA (blue) or AC (red) responses for each participant by their age in years 122
5.3 Distribution of response types for each age group 123
5.4 Mean AC (red) and SA (blue) scores for each adjective pair collapsed across age groups 125
5.5 Age in years of the youngest child to provide an AC response to each negative (solid red line) and positive (dashed blue line) adjective . . . 130
5.6 Number of unique responses to each negative (solid red line) and positive (dashed blue line) adjective 131
5.7 Number of unique responses to each negative (solid red line) and positive (dashed blue line) adjective, including the two-year-old's responses 132

6.1 Sample visual world display, for the sentence 'pick up the tall glass and put it below the pitcher', from Sedivy et al. (1999, p. 130, Fig. 4) . . 157

Chapter 1

Relative Adjective Semantics

1.1 Introduction

Relative adjectives, such as *big, tall* and *narrow*, are context dependent. The truth conditions of a sentence with an attributive relative adjective in the positive (bare) form, such as *tall* in (1), are dependent on the object being described, on a speaker's knowledge about properties of such objects and on the other objects in the immediate physical environment accompanying the utterance, among other things.

(1) Timmy is tall.

The sentence in (1) might be true of ten-year-old Timmy when uttered in his fourth grade classroom, but false when said in the gym as he plays basketball with a group of seventh graders. Formally, this variability is represented by interpreting such relative adjectives relative to a contextually determined comparison class. In order for Timmy to count as tall, for instance, he needs to have a degree of height that exceeds a standard of comparison[1] which is chosen relative to a comparison class,[2] e.g., the other children in the class or in the gym. The standard is determined by some function which ensures that the degree to which the object possesses the property stands out relative to the other objects in the comparison class (e.g., Bale, 2011;

[1]The standard of comparison is also known as a norm or a cut-off point (e.g., Tribushinina, 2011).
[2]The comparison class is also known as a reference class (e.g., Lassiter and Goodman, 2013).

Hare, 1952; Kennedy, 2007; Kennedy and McNally, 2005; Klein, 1980; Lassiter and Goodman, 2013; Siegel, 1979; Solt, 2011; Wheeler, 1972).

Language users must converge on similar representations of standards and comparison classes if they are to use and understand relative adjectives appropriately. Therefore, understanding the exact nature of a comparison class, what determines the composition of the class and what factors influence this determination is crucial to understanding the use and acquisition of dimensional adjectives. However, there have been few attempts to investigate these questions experimentally.

It is clear from the sentence in (1) that speakers use both world knowledge about the object being evaluated (Timmy in this case), as well as the immediate physical context in which the sentence is uttered, to determine the relevant comparison class. The speaker not only knows that Timmy is a ten-year-old boy and his height, but also knows things about ten-year-old boys in general and their height, about boys in general and their height, and about people in general and their height. This information, namely people's prior information about objects, object categories and their typical properties, may also play a role in determining the comparison class. Additionally, knowledge about the context of the utterance also seems to be relevant for determining the comparison class. For example, knowing that Timmy is a ten-year-old boy might, in some contexts, restrict the comparison class to contain only other ten-year-old boys. However, in other contexts, such as when (1) is uttered while Timmy is playing basketball with a group of seventh graders, even if the speaker knows Timmy's age, it might not be taken as relevant information, and the comparison class might contain the children playing basketball.

My main goal in this dissertation is to understand how information about the immediate physical context and world knowledge interact with each other in determining a comparison class and what the nature of the world knowledge involved is. In Chapters 2–4, I investigate factors from the context that are relevant in determining the comparison class, how they interact with a speaker's world knowledge about the objects in question and how the context influences the interplay between these factors. I investigate these questions experimentally, by manipulating the contexts in

which the adjectives appear. I examine how different contexts interact with speakers' prior world knowledge about the object categories in question to create different comparison classes which extend beyond the mere size of the objects on the screen. I find evidence to support the hypothesis that world knowledge is integrated with information from the immediate visual context to determine the comparison class. I differentiate between two types of world knowledge: knowledge of the real-world size of objects and knowledge about prototypicality properties of objects. I show that both of these types of knowledge are relevant in determining the comparison class, and I propose that features of the context, such as whether the images being judged are clipart-style images or photographs, are taken into account in this determination.

As a secondary goal, I am also interested in the acquisition process of relative adjectives. Despite speakers' different experiences of the world, they converge on fairly similar interpretations of these adjectives. I would like to understand how the context, and specifically the contrasts present in it, shapes the acquisition process. Relative adjectives (like other gradable adjectives) are often associated with antonyms, such as *big* with *small*, or *tall* with *short*. Learning these associations is part of acquiring the meaning of these adjectives. Probing children's antonym acquisition provides a window into their acquisition of these adjectives and insight into how they build up lexical entries in general. In Chapter 5, I present an experiment eliciting antonyms from children of a broad range of ages. The results of this experiment illuminate the process of acquisition of a semantic field of related adjectives and the roles context and contrast play in anchoring children's production of these adjectives.

To accomplish both of these goals, dimensional adjectives (e.g., *tall* or *wide*), which describe the physical extent of the nouns they modify, are taken as exemplars of the class of relative adjectives. Dimensional adjectives describe properties which are easy to represent visually and are thus easy to investigate. They are also a neatly contained set of adjectives, with each member having a clear antonym. In the studies in Chapters 2–4, which are concerned with the nature of the comparison class, its composition and the factors influencing it, I focus solely on *big* and *small*. This pair are the most general and first acquired dimensional adjectives (e.g., Tribushinina et al., 2014). They are also among the most common adjectives cross-linguistically,

even in languages with very restricted adjective classes (Dixon, 1982).

In the acquisition study in Chapter 5, I zoom out to explore the broader class of dimensional adjectives, looking at the gradual acquisition of this semantic field. In an experiment eliciting dimensional adjective antonyms in Hebrew from children aged 3;0[3] through 8;10, I show that the structure of the lexical domain shapes the acquisition process. I also show that the results of this experiment, conducted in Hebrew, are consistent with previous studies examining the acquisition of English dimensional adjectives (e.g., Clark, 1972) in terms of the order of their acquisition.

This dissertation examines how information from the immediate physical context is integrated with world knowledge when determining the comparison class for relative adjective interpretation and the nature of this world knowledge. In order to do so, we first need to situate ourselves in the semantics of relative adjectives, with a particular focus on the role of the comparison class and standard of comparison (Section 1.2) and the relevant theories of comparison classes (Section 1.3), as well as have a sense of the gaps in knowledge in this body of work (Section 1.4). Finally, as world knowledge in this context is often tied to object category representation, we should also situate ourselves in the literature on category representation (Section 1.5).

1.2 Relative Adjective Semantics

Relative adjectives are a subset of the class of gradable adjectives. Gradable adjectives have both semantic and syntactic identifying features. Semantically, they can be informally defined as "predicative expressions whose domains can be partially ordered according to some property that permits grading" (Kennedy, 1997, p. 1). For example, the domain of *tall* can be ordered by the measure of height and contains objects which are tall to varying degrees. Similarly, the domain of *dark* can be ordered by the measure of darkness and contains objects which are dark to varying degrees. However, non-gradable adjectives, such as *married* or *triangular* do not introduce the same type of ordering on their domains. While they do divide their domain into objects which are *married* vs. those which are not, or objects which are *triangular* vs. those which

[3]Throughout this dissertation, age is given in years;months.

are not, the objects in these domains are not *alive* or *triangular* to varying degrees (c.f. Lupyan, 2016, for a different perspective).

Syntactically, gradable adjectives have two characteristic properties. First, they can be modified by degree adverbials, such as *very* or *fairly* (Kennedy, 1997; Klein, 1980). Thus, sentences such as (2) are acceptable, whereas sentences such as (3) are not.

(2) a. Timmy is very tall.

b. My street is fairly dark at night.

(3) a. ?? Ben is a very married man.

b. ?? Fred's tent is fairly triangular.

Second, gradable adjectives can appear in degree constructions, such as comparatives, superlatives, equatives, *too* and *enough* constructions and *how* questions (Kennedy, 1997), as in the sentences in (4).

(4) a. Timmy is taller than Sam.

b. Timmy is the tallest boy in his class.

c. Timmy is as tall as Matt.

d. The living room is too dark to sew in.

e. The cat is thin enough to slip through the fence.

f. How long is your table?

Gradable adjectives can be divided into two main classes, absolute (non-context sensitive[4]), such as *empty* or *clean*, and relative (context sensitive), such as *tall* or *expensive* (Burnett, 2014; Kennedy, 2007; Kennedy and McNally, 2005; Lassiter and Goodman, 2013; Rotstein and Winter, 2004; Toledo and Sassoon, 2011; Unger, 1971). For example, in order to determine whether a cup is empty, it is not necessary to know anything about any other possible cups in the immediate vicinity, or other cups like the one in question. On the other hand, as shown for Timmy in (1), in order to

[4]Absolute adjectives may also be seen as less (but still somewhat) context sensitive than relative adjectives (McNally, 2011; Toledo and Sassoon, 2011).

determine if someone is tall, it is necessary to have information about the height of other people in the vicinity or typical heights of people in the same category.

The rest of this section will present one fairly standard formal semantic account of relative adjectives, with the goal of understanding how the notions of **standard of comparison** and **comparison class** have been integrated into such accounts. There are two main families of formal semantic accounts of the meaning of relative adjectives: degree-based accounts and non-degree-based accounts. In this dissertation, I use the language of the degree-based accounts to talk about adjectives, but the questions I deal with and the answers I provide are equally relevant to both accounts.

Adopting the degree-based account in Kennedy (2007), adjectives are measure functions, which are functions from individuals to degrees (type $\langle e, d \rangle$). Degrees are abstract representations of measurement, which are formalized as points or intervals partially ordered along some dimension (e.g., height). The set of ordered degrees constitutes a scale. Thus, *tall*, for example, is a function from individuals to degrees of height. The degrees are converted to properties of individuals via degree morphology, including *more, less, very* and *enough*, which introduces an individual argument for the measure function denoted by the adjective. Degree morphology imposes some requirement on the degree derived by applying the adjective to its argument, typically by relating it to another degree. For example, the denotation of *more* is shown in (5). *More* establishes an ordering relation between two degrees: one derived by applying the adjectival head to its external argument, the other by applying it to the constituent, marked by *than*. The *than* phrase is an argument that in comparative constructions introduces the standard of comparison, as in (4a).

(5) $[\![more]\!] = \lambda g \lambda y \lambda x . g(x) > g(y)$

Thus, the comparative predicate in (4a), *is taller than Sam*, is true of an individual if the degree to which that individual is tall is greater than the degree to which Sam is tall. This analysis derives the meaning of the bare positive form of the adjective, as in (1), from the meaning of the comparative. The bare positive form needs to somehow become a property of individuals but does not combine with overt degree morphology.

One way to derive the meaning of the positive form is to assume a null morpheme *pos*, which serves the same function as the degree morphology; it takes a measure function and returns a property of individuals, by relating the degree argument of the adjective to a standard of comparison (Bartsch and Vennemann, 1972; Cresswell, 1976; Kennedy, 1999, 2007; Von Stechow, 1984).[5] Various versions of the semantics of *pos* have been given.

In Kennedy and McNally (2005, ex. (13)) the denotation of *pos* is given as in (6), where g is the adjective and C is a variable over properties of individuals whose value is determined contextually, known as the comparison class.

(6) $[\![pos]\!] = \lambda g \lambda x . \exists d [\mathbf{standard}(d)(g)(C) \wedge g(d)(x)]$

On this account, the determination of the **standard** function is dependent on the lexical features of the adjective, specifically on its scale structure. In the case of relative adjectives, (6) requires a degree d to exceed a norm or average on the scale associated with g that is computed based on the comparison class identified by C. Other authors have proposed a variety of other options for what the function **standard** is, including the average (Bartsch and Vennemann, 1972), the top x percent or the top n items in C (Schmidt et al., 2009) or a degree which stands out in context relative to C (Kennedy, 2007).

While C is often implicit and must be inferred from context, as in (7a), it may also be made explicit linguistically, as in (7b), where it is expressed in a *for* phrase.

(7) a. John is tall.

b. John is tall for a jockey.

In the absence of an explicit comparison class, the adjective in (7a) is often interpreted as *tall for a man*, where a comparison class is recovered from the context based on what the speaker/hearer knows about John.

In the accounts of the semantics of relative adjectives just described, the standard of comparison and the comparison class play a crucial role in determining the interpretation. While various attempts have been made to make the standard of comparison fairly concrete, both in terms of theoretical proposals and experimental

[5]For an account in which the positive form is basic, see Kamp (1975).

support of these proposals (see Section 1.4), the comparison class has received less attention. The following section reviews various proposals regarding the nature of the comparison class.

1.3 Theoretical Framework for Comparison Classes

The interpretation of relative adjectives involves a comparison class along with a standard of comparison calculated using some function applied to this class. Thus, the exact placement of the standard of comparison is dependent on the composition of the comparison class. The theoretical literature remains fairly vague as to the nature of the comparison class. In this section, I present a roughly chronological overview of various proposals that have been made. In an early discussion of the meaning of the adjective *good*, Hare (1952, p. 133) expresses the intuition that different categories of objects have different standards for counting as *good*. Thus, "what distinguishes the class of comparison 'sports car' from the class 'family car' is the set of virtues which are to be looked for in the respective classes". Hare points out that the interpretation of *good car* will depend on what kind of car is under consideration. He recognizes the context dependency of adjectives in general, and of the comparison class (his class of comparison) more specifically, but does not attempt to formalize how the comparison class is determined, how it is represented cognitively, or how different kinds of contextual information are involved in determining its composition.

A more formal definition is provided by Katz (1972), whereby "the comparison class is the lowest-order category included as a semantic marker for the subject" (quoted by Siegel (1979, p. 244)). However, Siegel (1979) argues from the case of double measure adjectives (e.g., the distinction between *the fast heavy runner* and *the heavy fast runner*) that this is not accurate, as sometimes the comparison class is idiosyncratic. Instead, she proposes that the context of the utterance allows us to understand what exactly is included in the comparison class, but that there is no way to formally predict the set of possible comparison classes for a prenominal dimensional adjective, since adjectives that are not relevant to the dimension being measured may be omitted from consideration as in the case of *red-headed* in (8).

(8) Billy is a tall red-headed basketball player. (adapted from ex. (58) in Siegel (1979, p. 248))

Unlike the account in Katz (1972), Siegel's account recognizes that a given noun phrase might have multiple possible comparison classes, but leaves the determination of these classes to the context, without further specifying the process used. This is a desirable outcome, as evidenced in the example that opened this chapter, (1), where Timmy's height might be compared either to other children in his class, or other basketball players in the gym.

The rebuttal provided by Siegel (1979) is predated by a similar account by Bartsch and Vennemann (1972), who paraphrase *John is tall* as "John's degree of tallness is greater than the context specific or socially determined norm of a contextually determined comparison class", without specifying more about the contextual factors. Klein (1980, p. 13) is slightly more specific, defining a comparison class as "just the set of things that the participants in a conversation happen to be talking about at a given time. In formal terms, a comparison class is a subset of the universe of discourse which is picked out relative to a context of use." Similar notions have been formalized in later work, such as in Kennedy and McNally (2005, p. 349–350):

> "The standard of comparison is itself determined relative to a COMPARISON CLASS of objects that are similar in some way to whatever is being discussed (Klein, 1980); the result is that the truth conditions of sentences like those in [sentences similar to my (1)] may vary... *pos* encodes the relation **standard**, which holds of a degree d just in case it meets a standard of comparison for an adjective G with respect to a comparison class determined by $\mathbf{C}$, a variable over properties of individuals whose value is determined contextually... Since $\mathbf{C}$ is a free variable whose value must be fixed contextually... whether a particular degree exceeds the norm or not is a function of context. This allows for the sort of truth-conditional variability observed above."

In more recent, probabilistic accounts of adjective semantics (e.g., Lassiter and Goodman, 2013, 2017; Tessler et al., 2017), the comparison class (or reference class,

as it is sometimes called in this line of work) is treated as a statistical distribution of degrees which determines the value of the priors in the Bayesian model used for inference about adjective use. The value of the standard, or threshold θ_{adj}, is determined using probabilistic reasoning based on the background knowledge of the speaker about the distribution of the prior. The posterior probability of an interpretation in this model is derived by iteratively reasoning about the balance between a speaker's desire for informativity and a listener's beliefs about which utterances would be true. Thus, if the utterance in question were *Timmy is tall*, a weak interpretation where θ_{tall} is in the extreme lower region of the height prior is unlikely, since then the utterance would not be very informative. On the other hand, a strong interpretation where θ_{tall} is on the extreme upper region of the height prior is also unlikely, since that would raise by too much the likelihood of the utterance being false. This model provides a clear framework for integrating information from the context and world knowledge into the pragmatic reasoning involved in relative adjective interpretation. However, it leaves open the precise nature of the world knowledge, which information from the context is relevant and how different types of information are integrated into the comparison class.

Regardless of the theoretical framework chosen, all of these proposals require a context dependent standard determined relative to some contextually dependent comparison class. A considerable body of experimental work, reviewed in the following section, has probed the setting of the standard of comparison, but not much experimental work has looked at the determination of the comparison class. However, the interpretation of an adjective relies on both of these notions; without understanding the comparison class, it is difficult to interpret the standard of comparison, since it is dependent on the comparison class. Most of the work thus far does not attempt to probe the nature of the comparison class beyond the determination that it is in some way context dependent.

In this dissertation, the notion of context is split into two separate aspects: the immediate visual context and the speaker's world knowledge. I explore how these different kinds of contextual information interact and influence the comparison class

used to interpret an adjective. I adopt, and provide additional support for, the proposal that both of these notions of context are integrated in order to determine the comparison class. I further break down world knowledge into two types. I differentiate between knowledge about the real-world size of objects (Konkle, 2011) and knowledge about how we view the size of objects prototypically, e.g., that elephants are prototypically big (Tribushinina, 2011). I also hypothesize that properties of the visual context, such as whether the objects present are clipart images or photographs, influence the interaction between these two types of world knowledge in determining the comparison class.

As a foundation for understanding my experiments testing these hypotheses, the following section describes the previous experimental work that has been done on the interpretation of relative adjectives, most of which has focused on the standard of comparison, rather than on the comparison class. As these two notions are interdependent, investigations of each notion have repercussions for the other.

1.4 Previous Experimental Work

The semantics of relative adjectives involve a standard of comparison applied to a comparison class and both of these elements are context dependent, as discussed in Section 1.3. A significant amount of experimental work has been conducted in order to further flesh out these proposals. Much of this work has been focused on investigating the standard of comparison used to make adjectival judgments. Since the standard has been the main target of these investigations, the comparison class has, for the most part, been operationalized as the immediate visual context of the experiment. In many experiments, participants see a display containing images of multiple objects varying minimally in the degree of the property each has. Subjects are then asked to make adjectival judgments of various forms. The assumption implicitly made in interpreting the results of these experiments is that the objects in the display make up the entirety of the comparison class.

Given the theoretical work on comparison classes, it is clear that this assumption is an oversimplification. It is important to ensure that the factors influencing the

determination of the comparison class are understood in order to correctly interpret what the experimental results mean in terms of the standard of comparison. An alternate proposal has been put forth by Tribushinina (2011): the comparison class is actually determined by integrating information from the immediate visual context with world knowledge that a speaker has about the distribution of the adjectival property across objects of the category being described. This proposal forms the basis for the experiments I conducted, which are described in Chapters 2–4 of this dissertation. I extend this basic proposal and experimentally probe the interaction between various types of information in the immediate visual context and different types of world knowledge used in the determination of the comparison class.

In Section 1.4.1, I review select experiments focusing on the standard of comparison. I concentrate on how the notion of comparison class has been handled in these experiments and lay out the questions these experiments raise about comparison classes. In Section 1.4.2, I review the few experiments that have been conducted looking specifically at comparison classes. These experiments provide a foundation which I build upon in the remainder of this dissertation.

1.4.1 Experiments on Standard Setting

A number of experiments have been conducted examining the standard of comparison used in the interpretation of relative adjectives. In this section, I review three of these experiments. The first experiment I review is the first experiment in Solt and Gotzner (2012). They try to pinpoint the function used to set the standard as part of a broader program of investigating whether degrees are necessary in the underlying formal structure involved in the interpretation of relative adjectives. In this experiment, participants were shown displays, each with 72 eggs of varying sizes (e.g., Figure 1.1, from Solt and Gotzner (2012, p. 3, Fig. 1)). Four different distributions over four sizes of eggs were tested: Gaussian (largest number of eggs in medium sizes; fewer very small or very large); shallow Gaussian; linear (same number of eggs of each size); bimodal (2 peaks).

The participants were asked to indicate which eggs were *big* and which were *small*

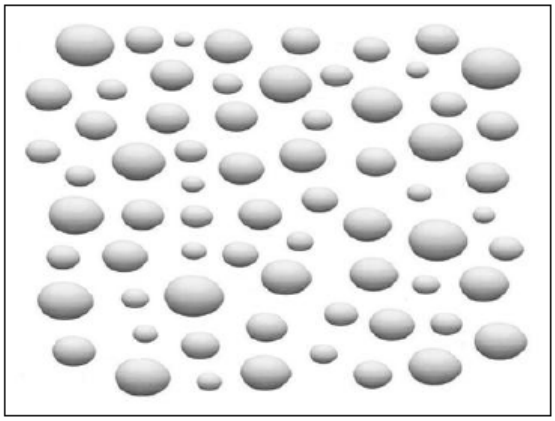


Figure 1.1: Sample stimulus set from Solt and Gotzner (2012, p. 3, Fig. 1)

in each display. The average number of eggs selected differed depending on the distribution shown. The authors interpreted this result as ruling out a function for determining the standard as simply picking the top $n\%$ of a ranking of members of the comparison class along the dimensions in question. Since each display consisted of the same number of eggs, the top $n\%$ would yield the same number of eggs selected as *big* or *small*, regardless of the distribution of sizes of eggs in the display. This conclusion implicitly relies on the premise that the display shown to participants makes up the entire comparison class.

Solt and Gotzner also found that there was no significant difference in the standards of comparison for *big* or *small* across conditions. Regardless of the distribution of the different sizes of eggs, there was no difference in the size of the biggest egg selected as *small* or the smallest egg selected as *big*. Therefore, they concluded that the standard might be calculated as a fixed proportion of the range of degrees possessed by members of the comparison class. In other words, an egg might count as *big* if its size is in the top $\frac{1}{3}$ of sizes for eggs in the comparison class. However, this conclusion also is only warranted if the comparison class in each trial is only the eggs in the display in that trial. If the eggs in the display inform the comparison class (accounting for the different number of items selected in each condition), but are not the entire comparison class, it is harder to rule out a standard of the top $n\%$ of eggs

(rather than sizes of eggs), since the exact composition of the comparison class in each case is unknown.

The second experiment I review is the first experiment in Schmidt et al. (2009), which is also focused on the function used to set the standard of dimensional adjectives such as *tall*. In this experiment the distributions shown to subjects were composed of novel objects (green rectangles of varying heights, assuming this experiment is the same one reported in Schmidt, 2009, p. 80) with labels like "pimwits". Since the pimwits were novel objects, subjects started with minimal assumptions about the distribution of heights in the comparison class. Each subject also only saw one distribution and therefore could not build up more information about the category over the course of the experiment.

In this case, the assumption that people are not integrating other prior knowledge with the information from the display in front of them to judge which pimwits are tall is more justified than in Solt and Gotzner (2012). However, the stimuli were simply green rectangles that look like bars on a graph that were given a label. There were no features indicating that they might have a function. People may well have some expectations about the size of pictures of rectangles or graphs. Therefore, while the distribution is likely to be heavily influenced by the immediate visual context of the experimental display, as it provides a large part of what the participants know about pimwits, there might still be some influence from participants' world knowledge. While the case of single trials involving novel objects is probably as close as possible to abstracting away from the effects of prior world knowledge, the question still remains how much influence other factors had on the comparison class.

In the third experiment I review, Hansen and Chemla (2017) approach the question of adjective interpretation from a slightly different angle. Their main question, which is less relevant to the present discussion, is whether color adjectives behave like absolute or relative adjectives. However, as controls for their investigation of color terms, they also include known relative adjectives (*tall, wide*) and minimum standard absolute adjectives (*spotted, dirty*) in their experiment (Experiment 2 in the paper, based on Syrett et al. (2010)). Each display showed two aliens, and participants were asked to "click on the [Adj] alien". The options "neither is!" and "both are!" were

also available (Figure 1.2 (Hansen and Chemla, 2017, p. 252, Figs. 5 and 6)). Each participant saw a total of 144, each contrasting two pictures of aliens.

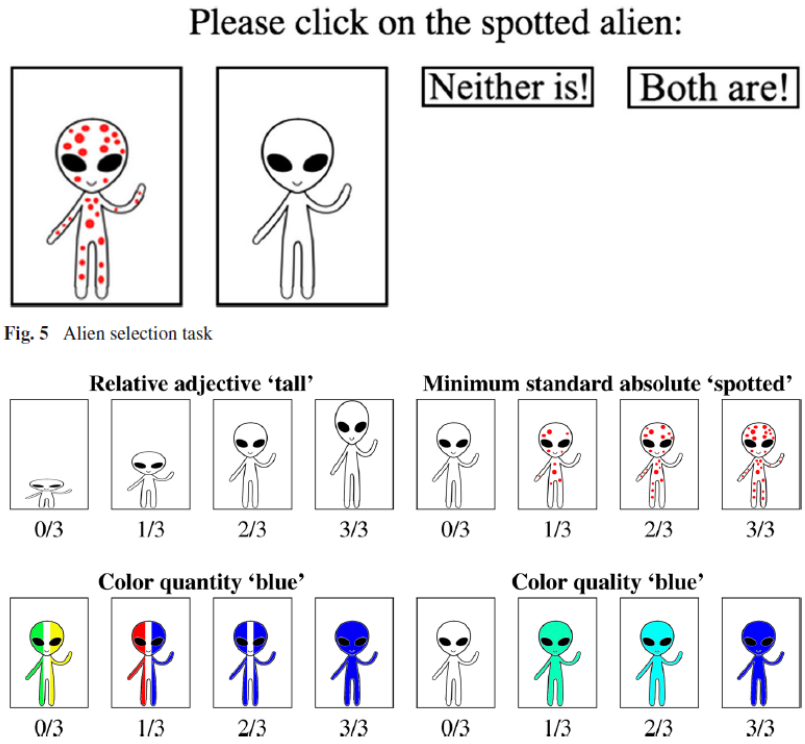


Figure 1.2: Stimuli from Hansen and Chemla (2017, p. 252, Figs. 5 and 6)

Based on the work in Syrett et al. (2010), Hansen and Chemla predicted that for relative adjectives, in each display, the alien with a greater degree of the relevant property would be selected. In explanation, Hansen and Chemla quote Syrett:

> Because relative GAs [gradable adjectives] such as *big* and *long* depend on the context for the standard of comparison, participants should posit a new standard of comparison each time a new pair is introduced in order to ensure that the adjective is true of just one object (i.e. the bigger or longer one). Thus, participants should always be able to accommodate the presuppositions of the definite description and accept the request as felicitous. (Syrett et al., 2010, p. 11)

However, this was not what Hansen and Chemla observed. When the display contained aliens 2/3 and 3/3 (see illustration in Figure 1.2), participants behaved as predicted and selected the taller of the two aliens. In contrast, when the aliens compared were 0/3 vs. 1/3 and 1/3 vs. 2/3, participants diverged from the predicted behavior and a substantial number answered "neither is!". The authors propose a number of different possible explanations of their results.

The first explanation considered is that the results are due to a threshold effect (Syrett et al., 2010; Kennedy, 2007) whereby there isn't enough of a difference in height between the aliens to call one *tall* but not the other. This is rejected since participants do behave as predicted when comparing 2/3 to 3/3, where the difference in height is even less extreme than in the conditions where participants diverge from the predicted behavior.

The second explanation proposed is that there is an order effect, such that as participants gain experience with these aliens over the course of the experiment, they become less willing to call the shorter ones *tall*. Hansen and Chemla also reject this explanation, since the "neither" response only becomes more probable over the course of the experiment for 0/3 vs. 1/3, but not for 1/3 vs. 2/3. They argue that if the order effect is the reason that participants are selecting the unpredicted "neither" responses, this should be the case for all of the trial types. Additionally, the order effect does not hold up when a correction for multiple comparisons is applied and also there seem to be order effects for other, non-relative adjectives as well. The authors take this to mean that participants are simply "performing the task differently by the end of the experiment across different types of adjectives" (Hansen and Chemla, 2017, p. 20).

The third explanation, which they accept, is that there is a lower threshold on the scale which must be surpassed in order for the adjective to behave like it has a contextually dependent standard. The example they give involves two tiny toy soldiers, one of which is noticeably taller than the other. Their intuition is that if asked to pass *the tall soldier*, one might object, arguing that neither is sufficiently tall. This solution is problematic. The lower threshold, as proposed by Hansen and Chemla, is not contextually dependent and is, therefore, the same for all objects.

However, given the context dependence of these adjectives, I would not expect this to be the case. While I share their intuition in the case of the soldier, I would predict that in the case of two grains of rice, one double the length of the other, there would be no problem referring to the longer one as *the long grain*.[6] In this example, one would think that this length would not be sufficient to exceed the lower threshold for *long*, but, since the object in question is a grain of rice, which is small by nature, the standard to count as *long* is fairly low.

Hansen and Chemla also raise the possibility of a geometric constraint (Barner and Snedeker, 2008), which imposes a requirement on the ratio between the horizontal and vertical dimensions in order to count as *tall* or *wide*. This is an interesting possibility and easily tested with other relative adjectives. While the two adjectives examined in this experiment may impose this restriction, others would not necessarily, particularly if non-dimensional relative adjectives were chosen. It would be interesting to see whether for non-dimensional relative adjectives, such as *dark* and *light*, where there is clearly no geometric constraint, results are similar to those found by Hansen and Chemla (2017), or more like those predicted by Syrett et al. (2010).

In light of the questions raised in Section 1.3, I would like to reconsider the option of order effects. The case of *tall* is particularly striking and, in absence of information about the stimuli for *wide*, it is difficult to know whether the same argument would hold for those as well. The same alien images were used as stimuli for all the adjectives in the experiment: absolute, relative and color terms. At the start of the experiment, after the instructions alone, participants had already seen 6 aliens, all of the same height (the introductory examples all related to color). The tallest alien is taller than all the other aliens, alien 2/3 is the same height as the rest (the initial six, as well as all of the aliens used for adjectives other than *tall*) and aliens 1/3 and 0/3 are shorter than all of the other ones. It may not take very many trials to decide that 0/3, 1/3 and 2/3 do not count as *tall* (and in fact the first two are short). Therefore, in order to argue that there are order effects, it may not be necessary for the probability of selecting "neither" to increase over the course of the experiment. It may be the case

[6]In fact, there is a type of rice known as *long grain rice*, which I would not think is long enough to exceed whatever lower threshold they propose.

that by the time the experiment starts, the participants have already begun to form a distribution of alien heights that is robust enough for them to not be willing to accept the existence presupposition of "the tall alien" at least some of the time (or for at least some of the participants).

The stimuli for *wide* are not shown in the paper, but if the same pattern holds with regard to the relationship between their width and the width of the aliens shown for the other adjectives, then I would propose that an actual increase in *neither is* responses over the course of the experiment is not necessary in order to support the order effects explanation. I would argue that participants may be building up a distribution of the dimensions of these aliens over the course of the experiment, starting from the three practice trials (where the aliens do not vary in height or width), and may set an expectation for "normal" alien dimensions. This distribution is then used over the course of the experiment to help determine the comparison class, despite there only being two aliens in the immediate visual context in each trial. One way to explore this possibility might be to use different aliens for each of the adjectives. If the only information participants have about the height of the aliens comes from the trials about height, then order effects might be more prominent.

It is interesting to note that participants' behavior is not consistent. Half of the responses for the comparison between 0/3 and 1/3 were as predicted, while the other half of the responses were that neither alien is tall, and about 60% of the responses for 1/3 vs. 2/3 were as predicted. The authors do not discuss how these responses distributed among participants and what the distribution might mean for individual variation on this matter. It might be the case that participants were consistent in their response strategy across adjectives, in which case some people may be more cooperative and willing to adopt the requisite existence presupposition. Therefore, by necessity, they might be willing to take the comparison class to be only the immediate visual context. The participants who responded that neither alien was tall may be more literal and unwilling to accommodate the existence presupposition. They may then use (the perhaps more natural) comparison class including all of the aliens they have seen in the experiment and choose the *neither is* option.

In the experiments discussed thus far investigating dimensional adjective standards, the concept of comparison class is operationalized as the immediate visual context. In Section 1.4.2, I review experiments probing the nature of the comparison class itself.

1.4.2 Comparison Class Experiments

While there has not been, to my knowledge, much experimental work focusing on how the comparison class is determined, there are two papers, taking very different approaches, that deal directly with this question.

Tessler et al. (2017) take a modeling perspective on the question of comparison class composition, looking at the relative utility of a general vs. more specific comparison class. Following the probabilistic approach, as described in 1.2, the comparison class is taken to be a statistical distribution of priors for the Bayesian Rational Speech Acts (RSA) model. Tessler et al. ask whether sentences with no explicit noun being modified by the adjective, such as *it's warm*, are more likely to be interpreted relative to a superordinate class (e.g., for a day of the year) or relative to a subordinate class (e.g., for a day in winter). They argue that a listener reasoning pragmatically about the interpretation of an utterance (a pragmatic listener) overall prefers subordinate comparison classes, though the extent to which the subordinate class is preferred is modulated by the likelihood that the adjective is true of the subordinate class. They argue that a subordinate class interpretation is more informative, since there is lower variance in the category (e.g., days in winter have less temperature variance than days of the year in general do).

There are two potential problems with this argument. The first problem is that the proposed model assumes explicitly that it is salient to the listener that the noun being modified is a member of the subordinate class (for details of the model, see Tessler et al., 2017, eqs. (4) and (7)). However, when speakers make a statement with no overt noun, such as *it's warm*, it may be the case that they are saying that it is not relevant what kind of day it is, or what time of year. The interpretation relative to the superordinate class may arise exactly when the type of day is not relevant;

therefore, building the relevance of the subordinate class into the model may not be appropriate. The experimental design was consistent with the model and made the membership in the subordinate class explicit and salient. Participants saw a prompt, e.g., *Tanya lives in Maryland and steps outside in winter*, followed by an adjective sentence such as *Tanya says to her friend "It's warm."* They were asked *What do you think Tanya meant?* and had to make a two-alternative forced choice between *It is warm relative to other days in winter* or *It is warm relative to other days of the year.*

Once the experiment explicitly mentions the season in the prompt and explicitly poses the option that the adjective in the sentence can be interpreted relative to days in winter, this interpretation is made salient, as is the day's membership in the subordinate class. These factors may conspire to increase the likelihood of selecting the subordinate class interpretation. In actual uses of such adjectives, the subordinate class membership may not be explicitly salient, possibly increasing the likelihood of a superordinate category comparison class interpretation.

The second potential problem is that Tessler et al. (2017) use higher informativity as one explanation of the preference for the subordinate class interpretation. However, it is also possible to reason differently about informativity. Tessler et al. (2017) reason that a subordinate comparison class interpretation is more informative since there is less variance in the comparison class. Therefore, establishing that the day is part of the subordinate class restricts the range of likely temperatures more. However, in another sense, the superordinate comparison class interpretation is also more informative. The utterance *it's warm* will be true in fewer cases, i.e. for fewer possible temperatures, if it is interpreted relative to all the days of the year, since the standard for being warm relative to any day of the year is higher. Taking a simplistic view in which for a day to count as warm the temperature must exceed the average temperature, if the average is calculated relative to the temperature throughout the year, the average will be higher than if it is only calculated relative to the winter temperature. Therefore, saying that it is warm relative to the superordinate class is more informative in the sense that the temperature must be higher – there are fewer possible temperature options. Changing the reasoning about informativity might change

the expected interpretation and thus, raise questions about the interpretation of the experiment.

Despite these issues, Tessler et al. (2017) provide an interesting approach to investigating the composition of a comparison class and its dependence on world knowledge. In these experiments, some of the vagueness in the interpretation of the sentence is due to there being no noun explicitly mentioned. Not only is the use of the adjective predicative (as opposed to attributive, as in *warm days*, see Bolinger (1967)), the subject of the sentence is a pronoun (*it*) without a clear referent. However, even when an explicit noun is present in the utterance,[7] the comparison class is still not fully determined, as pointed out by Siegel (1979) in response to Katz (1972).

Tribushinina (2011) addresses a case where the noun is explicitly mentioned, investigating the composition of the comparison class in predicative adjectives with explicit nouns as the subject of the sentence. She argues that speakers integrate world knowledge with the information coming from the visual context when determining whether an object counts as *big* or as *small*. In her experiment, subjects are presented with displays, each containing a set of seven stimuli pictures of a single object, e.g., seven elephants, or seven balloons, in a range of sizes, as in Figure 1.3 (Tribushinina, 2011, p. 539, Fig. 1). The sizes of the object shown (1–7 cm, increasing in 1 cm increments) do not vary across different objects. All of the images shown are smaller than the objects depicted in them. For each display, subjects are asked *Which [objects] do you find [big/small]?*. Each adjective appears with each object. Tribushinina compares the number of objects selected as *small* (the *small* zone) to the number of objects selected as *big* (the *big* zone). She posits that if the standard of comparison is skewed so that the *small* zone is bigger than the *big* zone, this indicates that knowledge of the size of these objects in the real world is integrated with the visual scale presented by the stimuli to determine the comparison class.

The objects in the immediate visual context were the same size regardless of which object appeared in the picture and of which adjective was heard in the prompt. If the comparison class was made up entirely of the objects in the immediate visual

[7]The noun may be made explicit either as the subject of a predicative adjective, as in (1), or as the argument of an attributive adjective, as in (8).



Figure 1.3: Sample stimulus from Tribushinina (2011, p. 539, Fig. 1)

context and the function used to set the standard is the same for both *big* and *small*, then the size of the zones should be equal. If the comparison class was determined using only world knowledge, based on prior experience with objects of the relevant category, then all of the images in the display would be considered *small*, since all of the images are smaller than the objects depicted. However, if world knowledge is integrated with information from the immediate visual context in order to determine the comparison class, then, Tribushinina reasons, the *small* zone should be bigger than the *big* zone. This integrates the knowledge that all of the images are in some sense small (when compared to the actual size of the objects depicted) together with the judgment relative to the scale presented, whereby the *big* and *small* zones should be equal. Across all object categories, the *small* zone was found to be bigger than the *big* zone.

In addition to the main overall effect of zone size, Tribushinina (2011) also found an effect of object type. The 12 stimuli objects were of three different size categories: prototypically big objects, like elephants; prototypically small objects like mice; and objects that were neither prototypically big nor prototypically small, like balloons.[8] She found that the *big* zone was bigger for prototypically small objects (e.g., mice), than for prototypically big objects (e.g., elephants). Conversely, the *small* zone was bigger for prototypically big objects (e.g., elephants), than for prototypically small objects (e.g., mice), as shown in Figure 1.4 (from Tribushinina, 2011, p. 541, Fig. 2).

[8]For a discussion of prototypicality status and its relation to actual size see Chapters 2 and 3, as well as Tribushinina (2011) and Tribushinina (2013a).

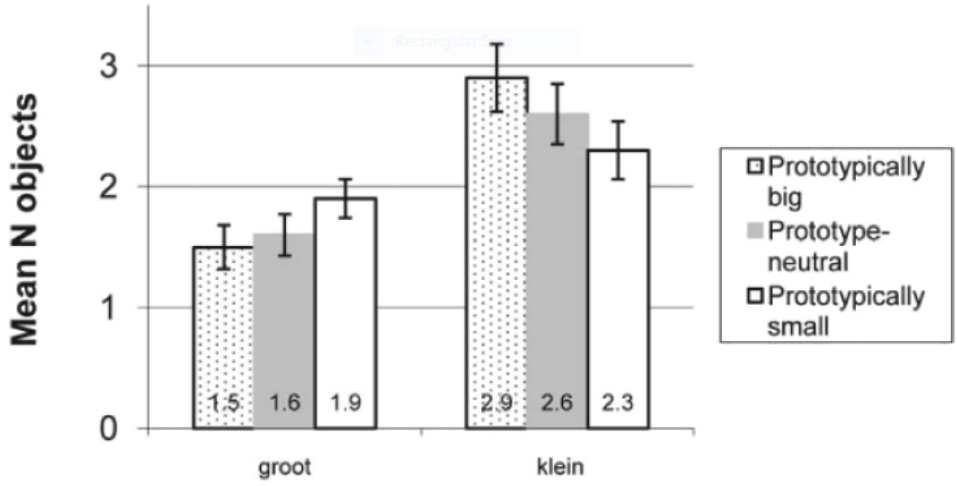


Figure 1.4: The mean number of images labeled *groot* ('big') and *klein* ('small') across prototypicality statuses, from Tribushinina (2011, p. 541, Fig. 2)

These prototypicality effects are taken as further evidence of integration of real-world size knowledge with information from the immediate visual context in determining the comparison class, as will be discussed at length in Chapter 3.

Tribushinina (2011) and Tessler et al. (2017) take the first steps in investigating the factors determining the composition of the comparison class and I build on this foundation in my work. While they provide initial evidence that speakers integrate world knowledge with information from the immediate context when determining the comparison class, much work remains to be done. I set out to further specify how this process of integration occurs and under what conditions. I investigate what kind of information falls under the label *world knowledge*, a crucial component in determining the comparison class under this approach.

Tribushinina (2011) operationalizes world knowledge as prototypicality status, as discussed in depth in Chapters 2 and 3. However, there is evidence that people store information about the actual size of objects in their mental representations (Konkle, 2011; Konkle and Oliva, 2011, 2012). Real-world size and prototypicality status do not always overlap. For example, there are prototypically small things, such as a baby, that in reality may be the same size, or in fact larger, than an

object which is considered neither prototypically small nor prototypically big, such as a balloon. Given that real-world size and prototypicality status are both stored in object representations but are separate things, there are a number of possible ways they might be utilized in adjective interpretation. One possibility is that only real-world size information is relevant, while another possibility is that only prototypicality status is relevant. A third possibility is that both prototypicality information as well as real-world size information are taken into account. One way the two types of information might be taken into account is that specific aspects of the context may encourage speakers to prioritize certain types of world knowledge over others, resulting in different judgments depending on the specific nature of the context. This dissertation takes an experimental perspective on world knowledge and begins to provide a more nuanced account of what this concept includes and the role it plays in dimensional adjective interpretation.

The relevant notion of world knowledge is intimately related to the mental representation of object categories. Knowledge about object categories is included in the world knowledge used in determining the comparison class for relative adjective interpretation. In (1), whether Timmy is or is not tall is judged based on, among other things, knowledge of heights of boys in the same category as Timmy. Part of determining the comparison class seems to involve determining the relevant category the object denoted by the modified noun belongs to. Often the comparison class is some subset of this category, where the context helps determine which subset. Thus, one of the factors in understanding comparison class representation is understanding category representation.

1.5 Categories, World Knowledge and Comparison Classes

This section reviews some of the relevant work on category representation in an effort to illuminate the world knowledge that speakers have regarding object categories, which in turn is critical to determining the comparison class. First, I will discuss

the work on categories and how world knowledge relates to category representation. Then I will lay out my hypotheses about how world knowledge regarding categories is involved in determining a comparison class.

Much work has been done on the nature of representations of concepts and categories (e.g., Barsalou, 1983, 1987; Casasanto and Lupyan, 2015; Hampton, 2006, 2007; Kumaran and McClelland, 2012; Margolis and Laurence, 1999; Medin and Schaffer, 1978; Nosofsky, 1987).[9] There are disputes about how stable people's representations are and how much is actually shared across speakers. The most radical position in this debate argues that all categories and lexical meanings are generated ad hoc in the context of use, based on contextual cues and the hearer's prior experience with those lexical items and categories (Casasanto and Lupyan, 2015). Under this view, interlocutors do not share as much meaning as it seems and never fully understand each other. Communication is successful when the participants succeed in their communicative goals. Only as much information as is needed for the accomplishment of the goal of the exchange must be conveyed. Under less radical views, while some categories are generated ad hoc (e.g., things on Jim's desk), others are more stable (e.g., furniture), with some consistent core of meaning (Barsalou, 1983, 1987).

Most of this work on categories has investigated the nominal domain. Comparison classes straddle the nominal and adjectival domains – on the one hand they are representations of nominal categories, but the relevant factor being considered is the degree to which these items possess a property which is lexicalized by an adjective.

An adjective's comparison class can be thought of as related to the mental representation of the category of the noun being modified by that adjective. It may include representations of the elements of some relevant subset of the exemplars of the nominal category in question which the speaker has encountered. I remain agnostic whether the comparison class only accesses the property in question as it pertains to the category of the object, or more complete representations. In other words, if the sentence *This elephant is big* is uttered, it may be the case that only the sizes of other elephants are accessed. Alternatively, more complete representations of other

[9]The terms concept and category have both been used in the literature. For the purposes of this dissertation I will use the term *category*.

elephants might be accessed, containing various properties, including the relevant property of size.

In either case, it seems plausible to posit that there is enough detail in the representations accessed to construct a probability distribution of the degree of the relevant property across the exemplars, i.e. instances previously experienced and stored, in the comparison class representation (Medin and Schaffer, 1978; Nosofsky, 1987; Kumaran and McClelland, 2012). Thus, to take the example of elephants discussed in Section 1.4.2, there is plausibly enough detail in the representations of various elephant exemplars previously encountered to construct a mental distribution of their sizes. This distribution might be thought of as the comparison class (e.g., Lassiter and Goodman, 2013). The standard of comparison is then calculated (in one of the ways discussed in Section 1.2) from this distribution.

In this dissertation, I argue that the comparison class takes into account world knowledge about the nominal category involved. The experiments in Chapter 3 explore the types of world knowledge involved in determining a comparison class. I propose two distinct types of world knowledge: information about the the distribution of possible sizes of objects of this category and information about whether the object category is taken to exemplify a prototypically big or prototypically small object (e.g., elephants are prototypically big objects, mice are prototypically small objects). This will be referred to throughout the dissertation as the prototypicality status of an object. These experiments also examine how the exact objects under consideration (e.g., whether the stimulus images are clipart-style images or photographs) influence the type of information taken into account in determining the comparison class. Using a paradigm adapted from Tribushinina (2011), I expand on the hypothesis that information from the immediate visual context is integrated with world knowledge to determine the comparison class. I posit that different types of world knowledge influence the listener, including knowledge of the prototypicality status (the size category the relevant object falls into) and knowledge of the distribution of actual sizes of the object. I argue that the relative influence of the different types of world knowledge utilized is affected by features of the context, including the type of depictions in the context, e.g., photographs or clipart-style images.

1.6 Roadmap

Language is rarely used in a vacuum; speakers bring a lot of knowledge of the objects in the world around them to bear when using language. One area where the interaction between language, speaker's world knowledge and the context of use is particularly salient is relative adjectives. The main goal of this dissertation is to examine the interaction of the physical context of an utterance and a speaker's world knowledge and investigate the nature of world knowledge in the context of relative adjective interpretation and acquisition. To accomplish these goals, I first describe a series of adult interpretation experiments which provide a more nuanced understanding of the notion of a comparison class than previous literature, both theoretical and experimental. Building on the idea that speakers' world knowledge influences their use and interpretation of relative adjectives, I seek to better understand what that world knowledge consists of and how different types of context may influence the world knowledge used in determining the comparison class. I then approach the issue of relative adjectives from the perspective of child language acquisition. I show that providing a contrastive visual context provides an anchor for children to produce relative adjective antonyms. I argue that the structured nature of the meaning of these adjectives shapes the gradual acquisition of the semantic field.

In Chapter 2, I present a norming study for the stimuli used in the adult interpretation experiments. In Chapter 3, I describe four experiments investigating the integration of world knowledge of various types with the immediate visual context in making adjectival judgments. In Chapter 4, I present two experiments which support the argument that the effects found in Chapter 3 are actually due to this integration of world knowledge and information from the immediate context, rather than to differences between the adjectives investigated.

In Chapter 5, I zoom out from looking at just the adjectives *big* and *small* and presents an acquisition study of dimensional adjectives. I propose that since context and comparison classes are crucial to the interpretation of these adjectives, providing clear contextual contrastive cues will make the retrieval and production of their antonyms easier, providing a more nuanced picture of the acquisition process than

was previously possible. Through an antonym elicitation study, I show the gradual structured process of acquisition of dimensional adjectives. In Chapter 6, I bring together all of these results and discuss their significance and contribution to the literature.

Chapter 2

The Size of Things: A norming study

2.1 Introduction

The first part of this dissertation contains a series of experiments which further investigate how world knowledge is integrated with information from the immediate visual context in determining the comparison class for relative adjective interpretation. These experiments use a paradigm and stimuli similar to Tribushinina (2011) (see Section 1.4.2), but in English rather than Dutch. Because of the change to English and because of the new stimuli, it is first necessary to have an understanding of how those stimuli are viewed by naive participants independent of my studies, to ensure that the results are comparable to those in Tribushinina (2011). Therefore, I ran a norming study looking at the perceived size of all of the stimuli used across Experiments 1–5 (presented in Chapters 3 and 4).

The experiments focus on the adjectives *big* and *small*. In all of these experiments participants see a display of seven images of the same object (e.g., elephants, mice, or beans) of a range of sizes, and are asked *Which [objects] are [big/small]?*. The goal is to provide support for the Integration Hypothesis, whereby world knowledge of size is integrated with the information in the immediate visual context when determining the comparison class for the adjective.

In Tribushinina (2011), the notion of world knowledge of size is operationalized as one form of prototypicality status. The intuition behind this is that one of the things that people know about elephants is that they are *big*, and that one of the things that they know about mice is that they are *small*. The stimuli are images of twelve objects, chosen such that there are representatives of three different categories of prototypicality status: prototypically big objects, prototypically small objects, and objects which are neither prototypically big nor prototypically small (henceforth known as neutral). Prototypicality is discussed at length in Section 3.1.

Tribushinina conducted her studies in Dutch, and assigned each object to a prototypicality status based on Dutch data, whereas I conducted my experiments in American English. It is plausibly the case that Dutch and American societies are culturally similar enough for the prototypicality statuses not to change. The primary goal of the norming study described in Section 2.3 is to ensure that this is in fact the case. A secondary goal of the study is to examine whether the two prototypically small stimuli that were changed between Experiments 1a and 1b, as described in Section 2.2.2, still behave like the other prototypically small stimuli. Section 2.2 describes the stimuli used in each of Experiments 1–5, followed by Section 2.3 which describes the norming study conducted on these stimuli.

2.2 The Stimuli

Experiments 1–5 all use a similar paradigm, but with different stimuli. Experiment 1a used the exact stimuli from Experiment 1 in Tribushinina (2011). These are described in Section 2.2.1. However, there are two problematic stimuli in this set, so in Experiment 1b, I replace them with two slightly different ones. The replacement stimuli are described in Section 2.2.2. Experiments 2 and 3 examine the effect of stimuli type on the comparison class, questioning the equivalence between prototypicality status and actual size in the real world. The stimuli in these experiments are described in 2.2.3. An alternate explanation might be proposed for the results of Experiments 1–3. Experiments 4 and 5 test predictions of this alternate explanation, and based on the results, it is rejected, providing further support for the Integration Hypothesis. The

stimuli for these last two experiments are described in Section 2.2.4. The description of the norming study itself follows in Section 2.3.

2.2.1 The Stimuli for Experiment 1a

Experiment 1a uses the exact stimuli from Tribushinina's study. The stimuli are clip-art style images of 12 objects, chosen to represent three categories of prototypicality status. The objects are listed in Table 2.1.

Small	Neutral	Big
mouse	balloon	elephant
gnome	umbrella	plane
chick	cake	hippo
baby	monkey	house

Table 2.1: Stimulus objects grouped by prototypicality status

These stimuli were chosen in a multi-step process described in Tribushinina (2013a). First, exemplars of bigness and smallness were selected from Dutch children's books teaching dimensional adjectives. Those exemplar objects had to be known and used by two-year-olds, as demonstrated using the Groningen Corpus (Bol, 1995). Additionally, the strength of association with bigness and smallness in child speech and child-directed speech was examined in the Dutch CHILDES corpus (Tribushinina, 2013b). For prototypically big objects, more than 70% of the size descriptions of those objects in the corpora had to involve the adjective *groot* ('big'), and similarly for the prototypically small objects and *klein* ('small'). Neutral objects were selected from objects described by *groot* and *klein* equally.

Finally, the list of objects generated by these criteria was judged by 15 adult Dutch speakers in a scaling task, where they were asked to indicate how big the objects were on a 10 point scale. Only the objects with a mean score above 7 were selected as prototypically big, and only those with a mean score below 3.5 were selected as prototypically small. The mean scores of objects in the neutral category were between 3.5 and 7. It is not made clear in the paper whether the adults were responding based

on the category labels, or whether those labels were also accompanied by an image, though I suspect the former. Additionally, the asymmetry in the cut-off points (3 points away from the end of the scale for *big*, but 3.5 points away for *small*) is not explained, and may have been chosen post-hoc based on the results.

2.2.2 The Stimuli for Experiment 1b

The stimuli in Experiment 1b are exactly the same clipart-style images as those in Experiment 1a, except for two substitutions. The chick is replaced with a duckling and the mouse is replaced with a bunny. The argument made in interpreting the results of Experiment 1a relies on all of the images being smaller than the objects depicted in them. However, the biggest images of the mouse and chick are close in size to an actual mouse and chick. Therefore, for the argument to go through, these two objects should be replaced with similar objects that are slightly bigger, but still considered prototypically small. One of the goals of the norming study described in Section 2.3 is to ensure that these new stimuli are also considered prototypically small.

2.2.3 The Stimuli for Experiments 2 and 3

The stimuli for Experiments 2 and 3 are photographic images of the same objects used in Experiment 1b. The photos are all close to square so that the aspect ratio of each image is maintained when it is scaled to different sizes in the experimental display.

2.2.4 The Stimuli for Experiments 4 and 5

The stimuli for Experiments 4 and 5 are photographic images of objects that are all very small. The photos are all bigger than the objects depicted in them, and are also close to square in shape. The objects used in these experiments are: ant, bean, blueberry, candy, earring,[1] fly, hazelnut, ladybug, pill, pin,[2] seed, and tack.

[1]The image used is of a diamond stud earring.
[2]The image used is of a pushpin.

2.3 Method

In order to norm the stimuli from Experiments 1–5, all of the objects were pooled together and combined with fillers in order to ensure an equal number of objects of each prototypicality status (prototypically big, prototypically small, and neutral). Subjects, recruited online using Amazon's Mechanical Turk (MTurk), were asked to rate the size of each object. The size data is then analyzed to examine the relationship between prototypicality status and size rating. Additionally the size ratings of the substituted stimuli from Experiment 1b (duckling and bunny) are compared to those of the objects they replaced from Experiment 1a (chick and mouse).

2.3.1 Participants

I recruited 134 participants using MTurk, 52 of whom were female, 80 male and 2 other. 5 participants were excluded for doing the experiment on screens smaller than 12 inches, and 2 were excluded for reporting a native language that was not English.

2.3.2 Materials and Procedure

Participants were arbitrarily assigned to one of three conditions. In all conditions the category label of the object was displayed (e.g., *elephant*). In the first condition the label was accompanied by a clipart-style image (like the stimuli from Experiments 1a–b), in the second condition it was accompanied by a photograph (like the stimuli from Experiments 2–5) and in the third condition the label appeared with no accompanying image. The clipart images for the tiny objects were created for this study and do not appear in any of the subsequent experiments.

In all conditions, participants were asked to rate the size of the object on a scale marked at one end with the label "sesame seed" and at the other end "roller coaster" using a slider bar (Figure 2.1). The labels were chosen so that the small end of the scale (sesame seed) was smaller than the smallest object used in any of the experiments, and the big end of the scale (roller coaster) was bigger than the biggest object used in any of the experiments.

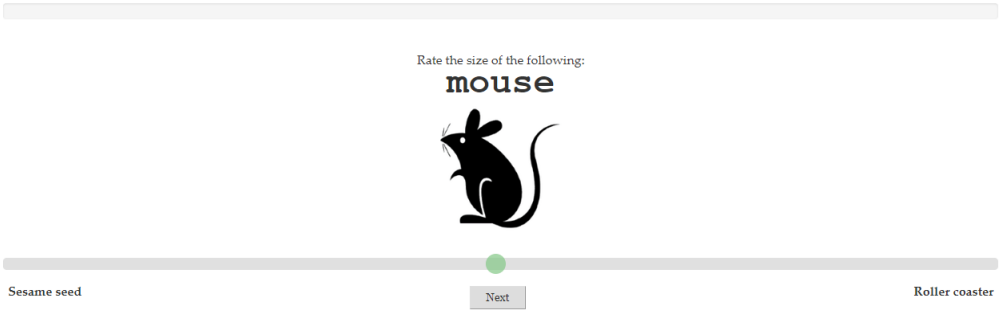


Figure 2.1: Sample trial display in the clipart condition, where the participant is asked to rate the size of a clipart mouse

The scale was coded as ranging from 1 to 100, but the numeric rating was not visible to the participants (see Figure 2.1). The objects rated included all 14 stimulus objects to be used in Experiments 1–3, the 12 tiny objects to be used in Experiments 4–5, and an additional 22 fillers, 6 small, 8 large, and 8 that are neither small nor large. Both bunnies and ducklings as well as mice and chicks were included. For a more in-depth discussion of the change from one to the other, see Section 3.4.1. The fillers were included to bring the number of items of each prototypicality status up to 12, determined by the 12 tiny objects (Table 2.2).[3] The filler objects fall into approximately three size categories, are easily depictable and include both animate and inanimate objects, similar to the target stimuli.

Participants saw all 48 objects, in randomized order. After the final trial participants were asked to provide demographic information such as gender and native language, as well as the size of the screen on which they performed the task.

2.3.3 Exclusions

Examining the raw data, there appear to be a large number of outliers, as seen in Figure 2.2. I calculated which trials deviated more than two standard deviations from

[3]There are 12 tiny objects so that Experiments 4 and 5 have the same number of items as the other experiments. Since all of the objects in Experiments 4 and 5 are smaller than the images depicting them, there are no prototypicality distinctions among the objects and, thus, there are 12 objects of the same category.

Tiny	Small	Neutral	Big
ant	**baby**	**balloon**	**elephant**
bean	**bunny**	**cake**	**hippo**
blueberry	**chick**	**monkey**	**house**
candy	**duckling**	**umbrella**	**plane**
earring	**gnome**	ball	bridge
fly	**mouse**	chair	bus
hazelnut	chipmunk	globe	car
ladybug	frog	goat	church
pill	hamster	kite	dinosaur
pin	puppy	microwave	giraffe
seed	rattle	sheep	piano
tack	teddy	typewriter	tree

Table 2.2: Stimuli for each prototypicality status. The bolded objects appeared in Experiments 1–5, the others were fillers.

the mean size rating of all of the objects in their prototypicality status category (big, small, neutral, tiny).

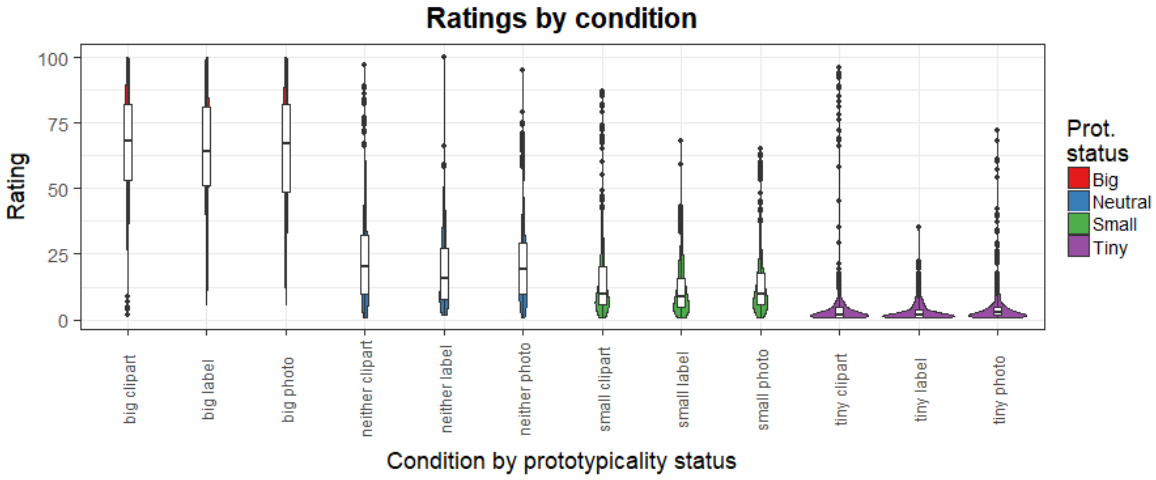


Figure 2.2: Violin plot of ratings by prototypicality status category in each condition

Examining the data of the eight participants with six or more deviant trials (out of 48), as shown in Figure 2.3, it seems that three of them are likely answering at random, since the distribution of their ratings for the different prototypicality status categories overlap almost entirely (anon-119, -32, -55 in Figure 2.3). It is clear that these participants should be excluded. Three participants seem to be using the scale very differently from others, but do not seem to just be answering at random (anon-127, -64, -79 in Figure 2.3). The curves for each prototypicality status category are distinct from the others. Another two participants grouped their ratings very far to the left of the scale (anon-22, -83 in Figure 2.3). Further work is required to understand how these participants were viewing the task.

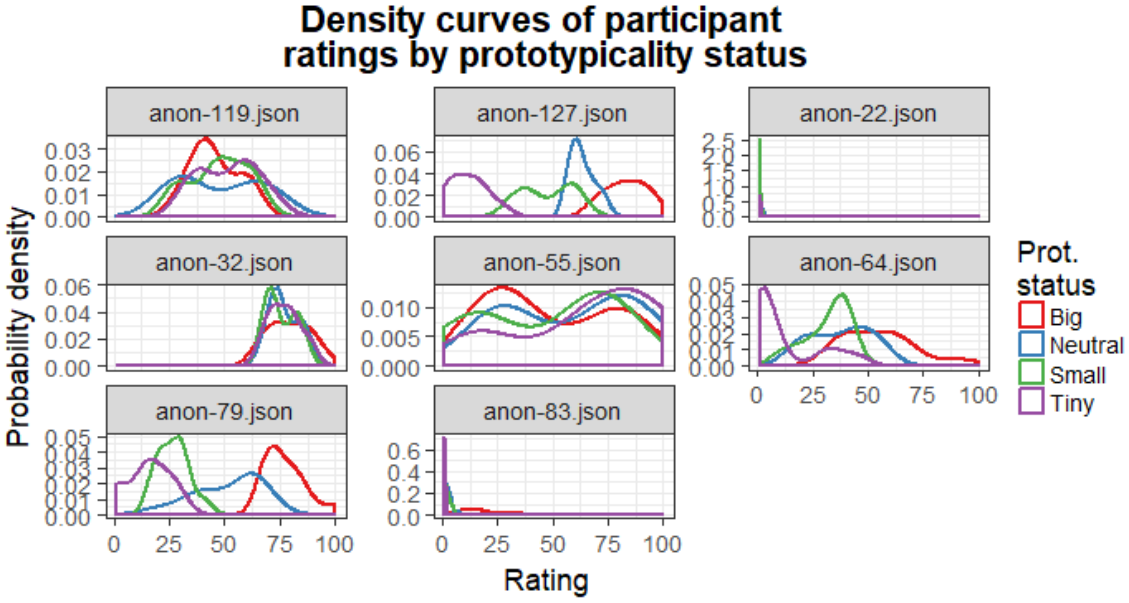


Figure 2.3: Density curves of ratings by participant for participants with 6 or more ratings that are more than 2 standard deviations from the overall mean rating of that prototypicality status category

For present purposes, all participants with six or more deviant responses were excluded (the participants in Figure 2.3), as were all of the deviant trials. After these exclusions, the dataset includes 5594 trials (of which 3053 are not fillers), for 119 participants split into 3 conditions. In condition 1, where participants saw photographs, n = 41, in condition 2, where participants saw clipart, n = 36, and in condition 3, where participants saw only the label, n = 42.[4] For comparison, Figure 2.4 shows the density curves of the responses of eight arbitrarily chosen participants, for most of whom the four prototypicality status categories are clearly distinguishable. Participant anon-39, for whom the categories are least distinguishable had five deviant trials, just below the threshold for exclusion.

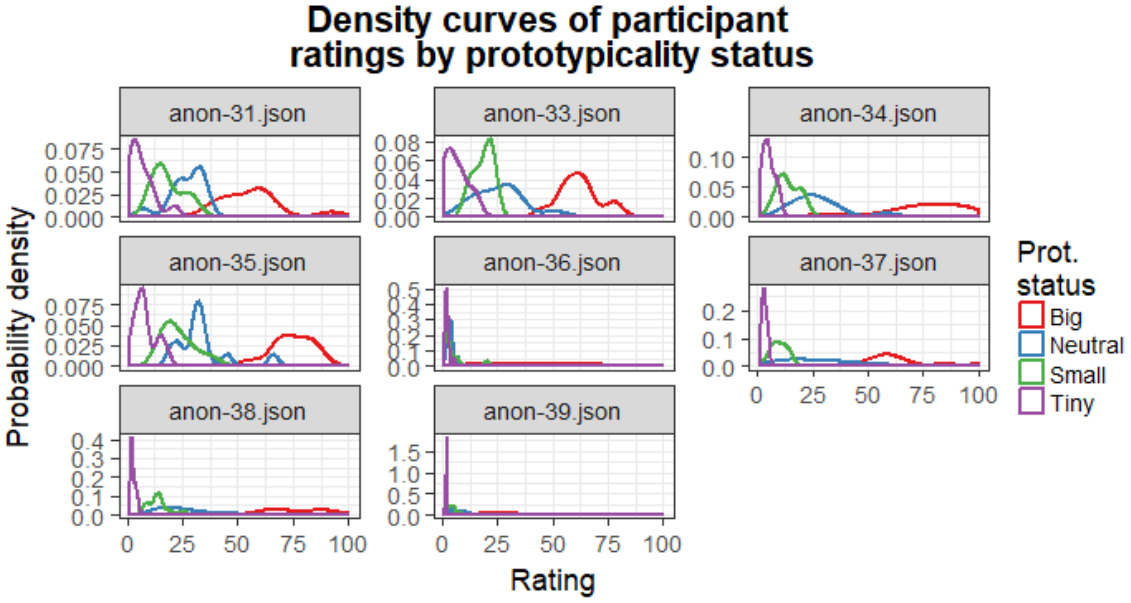


Figure 2.4: Density curves of ratings by participant for eight arbitrarily chosen participants who had fewer than six deviant trials

[4]Four of the excluded participants were in the photo condition and four were in the clipart condition. None were in the label only condition.

2.4 Results

As predicted, the prototypically big objects were rated as significantly larger than all of the other objects, and the tiny objects were rated as significantly smaller. Overall, the mean ratings of the different prototypicality categories are different from each other, as shown in Figure 2.5.

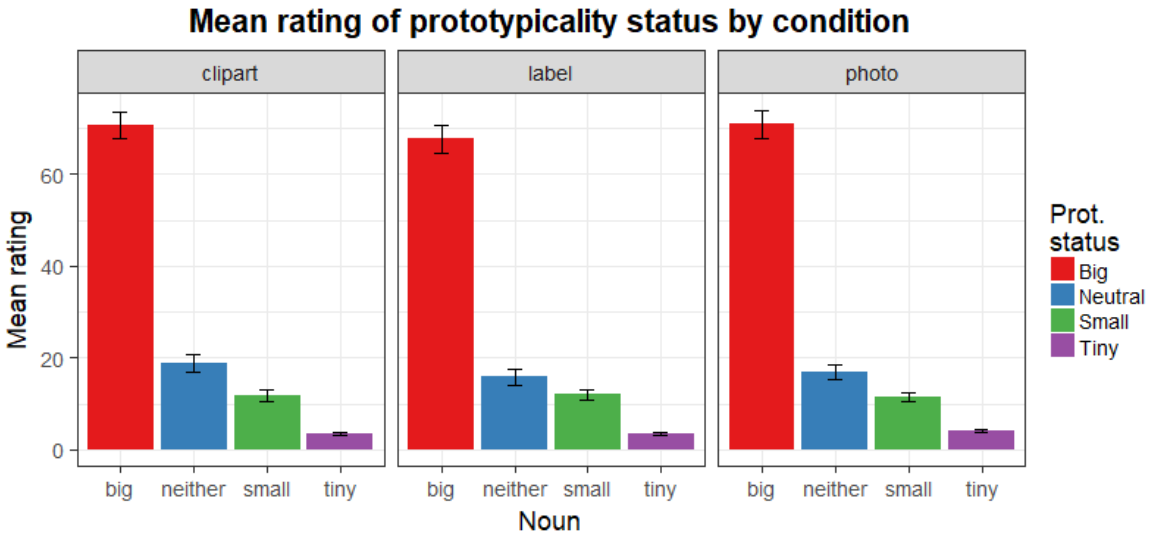


Figure 2.5: Mean ratings of the objects by prototypicality status in each condition. The error bars represent 95% confidence intervals (CIs).

When the results are broken down by object, the prototypically big and tiny objects are clearly distinct from the others, but it seems that there might be item effects for the prototypically small and neutral objects. The two categories seem to have some overlap, as seen in Figure 2.6.

In order to examine the effect of prototypicality status and stimulus type (condition) on the size ratings, the data were analyzed using linear mixed-effects regression models with planned orthogonal contrasts calculated using the lme4 package (v. 1.1-15) (Bates et al., 2015) and the lmerTest package (v. 3.0-1) (Kuznetsova et al., 2017) in R (v. 3.3.3) (R Core Team, 2017). The contrasts were set up so as to compare the tiny objects to the neutral and small, the big to all of the rest, and the neutral to the

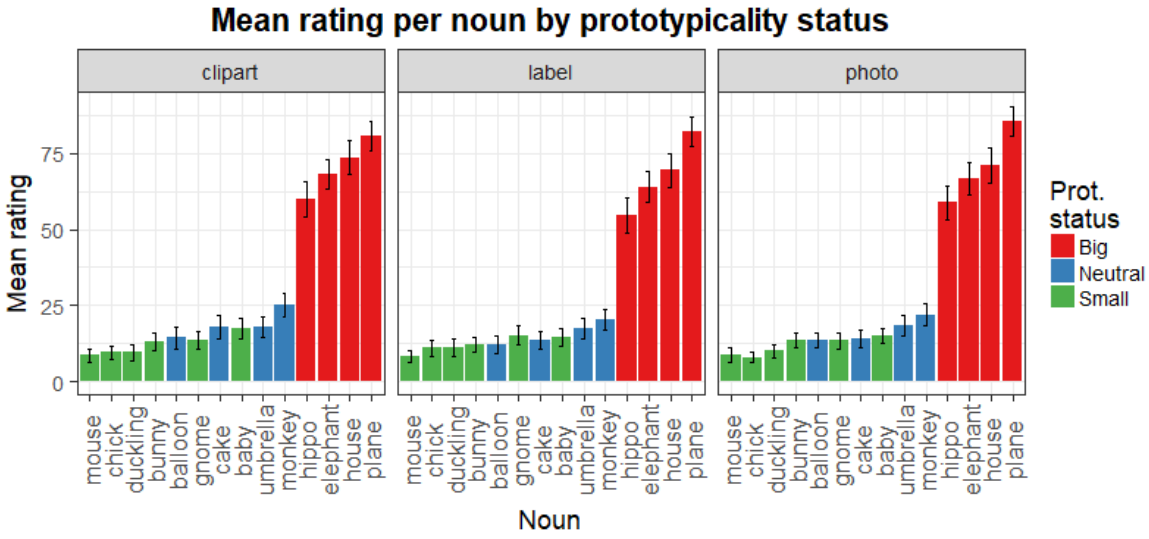


Figure 2.6: Mean ratings of the objects by noun in each condition. The error bars represent 95% CIs.

small. Maximal random effects justified by the design were used (Barr et al., 2013). The slope for each subject varied by prototypicality status, since each subject saw all of the objects, but did not vary for stimulus type, since each subject only saw one condition. The slope for object varied by stimulus type, since each object appeared in all three conditions, but did not vary by prototypicality status, since each object has a given prototypicality status.

The difference between tiny objects and the mean of neutral and small objects is significant($\beta = -7.75, p < 0.001$), as is the difference between big objects and the rest ($\beta = 43.22, p < 0.001$), and the difference between neutral and small objects ($\beta = 3.42, p = 0.011$). There is no significant effect of stimulus type (with no label being the baseline, there is no significant difference in size rating when the stimulus was a photograph ($\beta = 0.802, p = 0.138$), or when it was clipart ($\beta = -0.05, p = 0.941$)).

In a model including an interaction between prototypicality status and stimulus type, the picture changes somewhat. Prototypicality status was removed from the random effect term for subject in order to achieve model convergence. There is again a significant difference between tiny objects and the mean of neutral and small

objects ($\beta = -7.0016, p < 0.001$) and a significant difference between big objects and the rest ($\beta = 42.69, p < 0.001$). However, the difference between small and neutral objects is not significant ($\beta = 1.91, p = 0.197$). There is no significant main effect of stimulus type, but there are significant interactions between stimulus type and some of the other comparisons. The interaction between the comparison between big objects and all of the other objects and the stimulus type being photographs is significant ($\beta = 1.87, p = 0.008$). The interaction between the comparison between tiny objects and the mean of small and neutral objects and the stimulus type being clipart is marginally significant ($\beta = -1.006, p = 0.072$), as is the interaction between the comparison between big objects and the mean of all of the rest and the stimuli being clipart ($\beta = 1.57, p = 0.052$). The interaction between the comparison between neutral and small objects and the stimuli being clipart is significant ($\beta = 1.515, p = 0.016$). These results are summarized in Table 2.3.

Effect	β	p[5]
tiny vs. mean of neutral and small	-7.0016	< 0.001 ***
big vs. rest	42.69	< 0.001 ***
small vs. neutral	1.91	0.197
photograph	0.98	0.361
clipart	1.45	0.199
tiny vs. mean of neutral and small : photo	0.21	0.671
big vs. rest : photograph	1.87	0.008 **
small vs. neutral : photograph	0.85	0.124
tiny vs. mean of neutral and small : clipart	-1.006	0.072 .
big vs. rest : clipart	1.57	0.052 .
small vs. neutral : clipart	1.515	0.016 *

Table 2.3: Results for model with interaction between prototypicality status and stimulus type

[5] *** indicates significance at the $< .001$ level, ** $p < .01$, * $p < .05$, . $p < .1$.

Model comparison using ANOVA shows that the model with the interaction between prototypicality status and stimulus type accounts for more of the variance ($\chi^2 = 18.617, df = 6, p = 0.0049$).[6]

The secondary goal of this study is to evaluate whether, in Experiment 2, bunnies can be substituted instead of mice and ducklings instead of chicks. The goal of this substitution is to ensure that all of the images used are smaller than the objects depicted in them. However, the change makes comparing the results of Experiment 1 and 2 slightly problematic, as discussed in Section 3.4.1. Comparing the norming data for these pairs of objects provides a metric for evaluating the substitution. The average ratings for these four objects are shown in Figure 2.7.

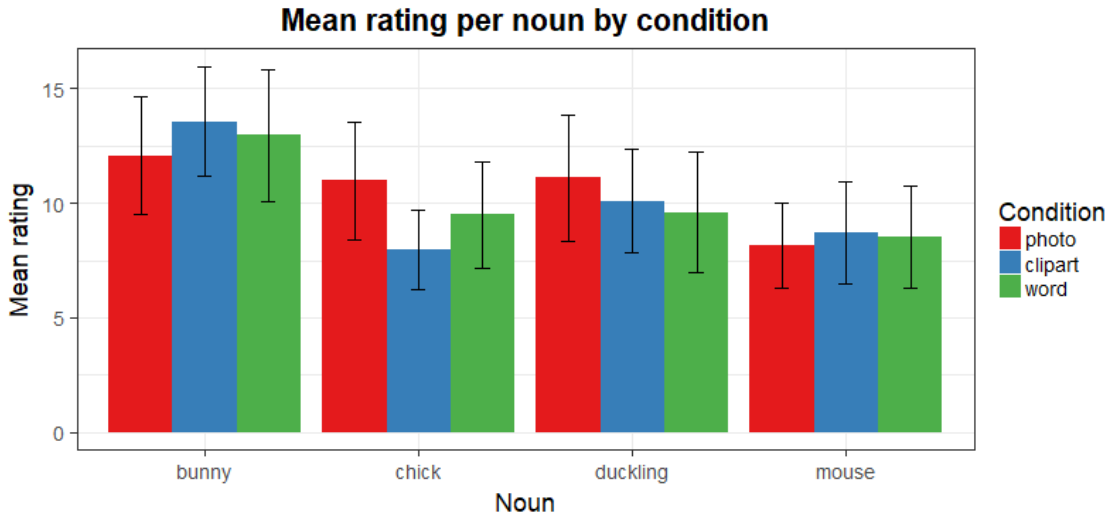


Figure 2.7: Mean ratings of bunny, mouse, chick and duckling by noun in each condition. The error bars represent 95% CIs.

The chick and the duckling are rated as fairly similar in size. However, the bunny is rated as significantly larger than the mouse in all conditions. When the mice and chicks are excluded, the overall category means remain fairly distinct (Figure 2.8).

[6]For the purposes of the model comparison an additive model was constructed with a reduced random effects structure to be comparable to the interaction model.

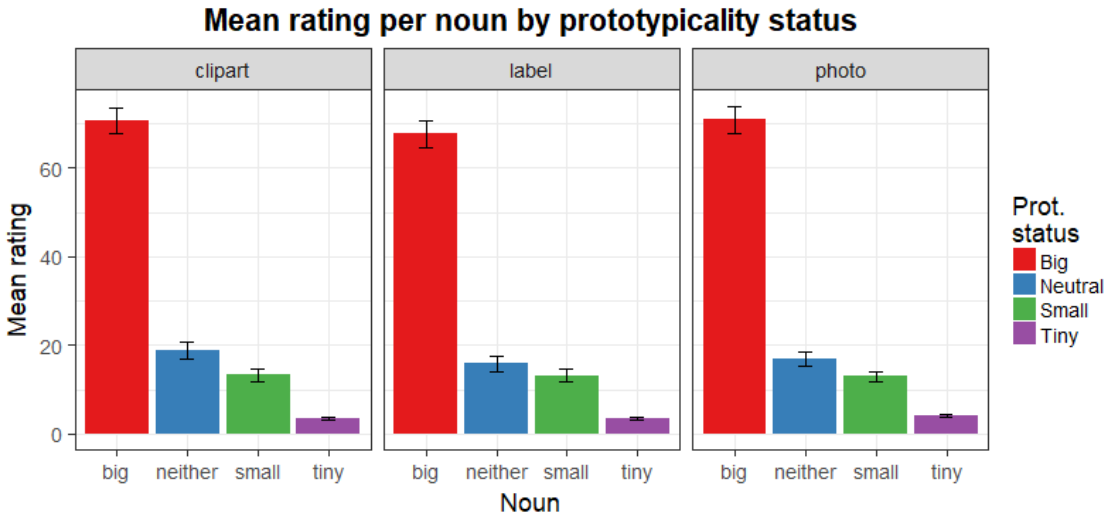


Figure 2.8: Mean ratings of the objects by prototypicality status in each condition for target objects, excluding mice and chicks. The error bars represent 95% CIs.

2.5 Discussion

The results of the norming study indicate that people perceive the size of the prototypically big objects to be significantly bigger than the other objects in the experiment. The tiny objects used in Experiments 4 and 5 are similarly perceived to be significantly smaller than the rest of the objects. The model accounting for the most variance, with an interaction between prototypicality status and stimulus type (condition), does not reveal a significant difference between small and neutral objects. However, there is an interaction between condition and this comparison. When the stimuli are clipart, there is a significant difference between prototypically small and neutral objects. This is in close alignment with the results of Experiments 1 and 2, as will be discussed in Section 3.6.

This task can also be used to evaluate the substitution of bunnies and ducklings for mice and chicks in Experiment 1b. I found that while chicks and ducklings seem to not be rated as being different sizes, bunnies were rated as significantly larger than mice. At first glance, this result may not bode well for the proposed substitutions. However, it may actually indicate that the ratings fairly closely reflect the real-world

size of objects, as bunnies are indeed larger than mice, rather than reflect knowledge about prototypicality status.[7] This hypothesis is consistent with the precise wording of the task, where participants were asked to rate the size of the object. It is possible that if the norming task were more precisely geared at accessing intuitions about prototypicality status, then bunnies and mice would behave more similarly. To get at this, the question might be phrased as "Which term would you use to describe the following:" with the response options being: "big", "small", and "neither". I predict that in such a task, the bunnies and mice and chicks and ducklings would be treated identically, all labeled *small*. This task would tap more directly into intuitions about prototypicality specifically, rather than about size in general.

It is promising for the substitution, however, that when the mice and chicks are excluded, the category means still remain fairly distinct, as seen in Figure 2.8. This is one indication that substituting bunnies for mice and ducklings for chicks will not change the prototypicality status of the stimuli, and therefore should not influence which images participants choose as *big* and *small* in Experiment 1b.

Overall, the results of the norming study support the classification of objects into prototypicality statuses, as in Tribushinina (2011) and Tribushinina (2013a). There are clear distinctions between all of the prototypicality status categories, except the prototypically small and neutral objects. The distinction between prototypically small and neutral objects is dependent on the stimuli type, and is significant only when the stimuli are clipart. This interaction will be important for the discussion in Chapter 3, comparing the comparison class results for stimuli which are clipart (Experiment 1a and 1b) and photographs (Experiments 2 and 3).

[7]This distinction between prototypicality status and real-world size is discussed at length in Chapter 3.

Chapter 3

How World Knowledge Shifts Adjective Interpretation

3.1 Introduction

As discussed in Chapter 1, the interpretation of dimensional adjectives like *big*, *small*, *tall* and *short* is dependent on a contextually determined comparison class. However, the exact nature of a comparison class and the way in which it is contingent on the properties of the objects the adjective is being applied to have not yet been adequately explored. In this chapter, I argue that speakers utilize both the objects in the immediate visual context and prior world knowledge about the relevant category of object in order to determine the comparison class, and therefore interpret a relative adjective. Moreover, world knowledge is not a single, monolithic body of stored information. Rather, world knowledge is comprised of different types of information. I distinguish between two different types of information: information about the real-world size of objects and information about the prototypicality status of objects (whether an object is considered, for example, to exemplify the property of being big). I argue that the relative influence of each type of information varies depending on features of the immediate visual context.

A comparison class (Cresswell, 1976; Kennedy, 1999; Klein, 1980) can be thought of as an underlying distribution of degrees of the property denoted by the adjective,

within the category of the object modified (Lassiter and Goodman, 2013, 2017). For example, *big* relates objects to their degree of size. In order to determine which elephants count as *big*, it is necessary to ascertain what the distribution of the property (size) is among members of the comparison class (relevant elephants). Then some function determines the standard of comparison, i.e. the degree of size which an object must surpass in order to count as *big*. As discussed in Section 1.2, one proposal is that the standard is determined by the population mean, so that to be *a big elephant*, the elephant must exceed the mean size of all elephants (Bierwisch, 1989). Various experimental work has been conducted regarding the formula involved in determining the standard of comparison, as discussed in Section 1.4.1.

Research on the function used to determine the standard of comparison provides information about one component of relative adjective interpretation, but this is not the only relevant component. The interpretation of relative adjectives is also dependent on the comparison class to which the standard is applied. As people experience the world around them and interact with objects and their depictions, they build up mental representations of distributions for various properties of those objects, including their size. These representations contribute to the comparison class used in interpreting relative adjective meaning, as discussed in Section 1.5. Thus, when encountering a new token of an object, its size can be compared to the distribution of sizes of relevant similar objects to determine whether its size exceeds the standard for counting as, e.g., *big*.

To date, much of the experimental literature has focused on the standard used for relative adjectives, and in doing so, has operationalized the comparison class as the immediate visual context in the experimental display. We know that in the real world comparison classes are more complex, but investigating this complexity is not simple. Initial support for the idea that other information is also used when determining the comparison class is found in Tribushinina (2011), described in Section 1.4.2. As discussed there, she argues for the integration of real-world size knowledge with information from the immediate visual context. She supports this argument with an experiment in which knowledge of the real-world size of objects is functionally equated with knowledge about their prototypicality status (whether the objects are

classified as prototypical instances of *big* objects, *small* objects or neither).

The notion of prototypicality Tribushinina uses is based on work such as Dirven and Taylor (1988). Although prototypicality is usually discussed in the nominal domain (e.g., Mervis and Rosch, 1981; Rosch, 1973, 1975, 1977), Dirven and Taylor import this notion into the adjectival domain. In the nominal domain, for example, people find a robin to be a better example of the category BIRD than an ostrich. While both robins and ostriches are members of the category BIRD, robins are more prototypical instances of the category. Rosch and colleagues have shown that subjects are quicker to identify prototypical instances as members of the category and are more likely to list them when asked to name instances of a category.

Dirven and Taylor (1988) suggest that, similarly, some nominal categories are more prototypical instances of the property TALLNESS than others. Men and trees might be considered prototypical instances of the property TALLNESS, while windows might not. Just as in the nominal domain it is easier to come up with a robin than an ostrich when thinking of the category BIRD, in the adjectival domain it is easier to conceptualize what is meant when the property is attributed to the nominal categories that are more prototypical instances of the property than to nominal categories that are less so. Thus, it is easier to conceptualize what is meant by, for instance, *tall man* or *tall tree* than *tall window*. Nominal categories less protoypically associated with the property might have lower frequency of collocation with the adjective, and lower acceptability ratings of the collocation. Dirven and Taylor (1988) also find that some nouns whose referents are ranked as prototypically tall, like *skyscraper*, are ranked low in acceptability in the collocation with *tall*, e.g., *tall skyscraper*. They propose that because SKYSCRAPER as a category is an optimal instance of TALLNESS, speakers do not tend to talk about *tall skyscrapers*, since it is redundant to attribute TALLNESS to any particular member of the category.[1] Prototypicality, as described by Dirven

[1]Higher frequency of collocation between an adjective and a noun that is a protoypical instance of the property associated with it is especially noticeable in child directed speech and literature directed at children. Young children are in the process of building up these networks of associations and forming category representations. When adults speak to other adults, these patterns are less noticeable since they are taken for granted and therefore, perhaps, not commented upon as often (Tribushinina, 2011). Thus, the collocation *big elephant* might be relatively frequent in child directed speech, but rarer in adult speech. The prediction that arises from Dirven and Taylor (1988) is that

and Taylor (1988), is also part of a speaker's knowledge about object categories, and thus might influence the composition of a comparison class.

Tribushinina (2011) uses the notion of adjectival prototypicality adopted from Dirven and Taylor (1988) to determine the prototypicality status of her stimuli, i.e. whether the objects in the experiment are considered prototypically big, prototypically small, or neither. She shows that the information regarding the prototypicality status classification is integrated with the size information from the immediate visual context to determine the comparison class. This result is interpreted to mean that real-world size information is integrated with information from the visual context in determining the comparison class. In doing so, the tacit assumption is made that the notion of prototypicality status reflects real-world size, or perhaps is the way real-world size is encoded.

The equation on Tribushinina's part of real-world size and prototypicality status raises an important question. The actual size of an object is not always consistent with its prototypicality status. For instance, while a baby is a prototypically small object, and a balloon is neither prototypically big nor prototypically small, a baby might be bigger than a balloon in many cases. Additionally, various studies by Konkle and colleagues (e.g., Konkle, 2011; Konkle and Caramazza, 2013; Konkle and Oliva, 2011, 2012) have shown that real-world size information is stored in some detail in object representations and affects the processing of images and visual scenes. Therefore, I propose that prototypicality status is a distinct notion from real-world size information.

Both prototypicality status and real-world size information play a role in determining the comparison class used for relative adjective interpretation. Cases where the two provide contradictory classifications, such as in the case of the baby vs. the balloon, may allow us to disentangle the affects of the two types of information. When the results for prototypically neutral objects, such as balloons, are indistinguishable from the results for prototypically small objects, such as babies, then I propose that real-world size information is playing a larger role in determining the comparison class. When the results for prototypically neutral objects differ from those of prototypically

this collocation would also be given a relatively low acceptability rating.

small objects, I propose that prototypicality status is playing a larger role.

If real-world size information and prototypicality are distinct types of world knowledge, then the results in Tribushinina (2011) actually demonstrate that prototypicality status affects the comparison class of relative adjectives. This is in line with the larger argument that world knowledge is integrated with information from the immediate visual context to determine the comparison class, but the original argument is imprecise.

The goal of this chapter is twofold. First, I aim to further the argument put forth in Tribushinina (2011) that world knowledge is integrated with information from the immediate visual context in determining the comparison class used for relative adjective interpretation. Second, I aim to show that world knowledge is multi-faceted and distinguish between two types of world knowledge: knowledge of prototypicality status and knowledge of real-world size . I propose that in different contexts one type of world knowledge will be prioritized over the other in determining the comparison class. Therefore, manipulation of the context will affect the comparison class used in relative adjective interpretation. The context of the utterance in the case of these experiments consists of displays of pictures of objects. Manipulation of these pictures will thus change features of the context.

The stimuli used in the study in Tribushinina (2011) were clipart images. Clipart images emphasize salient, prototypical features, and may vary widely from the actual appearance of the objects depicted. Therefore, they may trigger access to prototypical, generalized information about object categories, rather than information about the actual distribution of properties found among these objects in the world. The ease of access to prototypical information may, in turn, encourage the use of knowledge regarding prototypicality status in determining the comparison class.

I hypothesize that, when the stimuli are photographs rather than clipart, the effect of prototypicality status might be diminished. Photographs depict actual exemplars of objects, which have actual dimensions and proportions. Therefore, if the stimuli used are photographs rather than clipart images, real-world size might have a stronger effect than prototypicality status on the determination of the comparison class. This

effect will be clearest in cases in which prototypicality status and real-world size information diverge, as in the example of the baby and the balloon. I predict that when the stimuli are photographs rather than clipart, the results for prototypically neutral objects, such as balloons, will not be distinct from the results for prototypically small objects, such as babies.

Before I manipulate the context, I replicate the results from Tribushinina (2011) in an experiment conducted on Amazon's Mechanical Turk (MTurk) platform (Buhrmester et al., 2011) rather than in a laboratory setting, and I use English rather than Dutch. The stimuli in this experiment, as in the study in Tribushinina (2011), are clipart images. The results are expected to distinguish between objects of different prototypicality statuses, as in the original experiment. In Experiment 1b, I change the stimuli slightly to ensure that all images are actually smaller than the objects they depict, a key assumption in the reasoning about the expected results. In Experiment 2, I investigate the effect of stimulus choice on the comparison class. I provide evidence that the comparison class is composed differently when the objects being modified by the adjective are clipart images rather than when they are photographs, resulting in different size judgments. In Experiment 3, I examine the effect of the ordering of the stimuli on the response strategy and ensures that the results of the previous experiments were not unduly influenced by strategy.

A finding of different comparison classes for prototypically small and neutral objects in Experiments 1 vs. similar comparison classes in Experiments 2 and 3 demonstrates differentiation between knowledge of prototypicality status and knowledge about real-world size. Overall differences between the comparison class for *big* and the comparison class for *small* across all three experiments demonstrate that world knowledge is integrated with information from the immediate visual context to form these comparison classes. I argue that in Experiments 1a and 1b the world knowledge primarily relied on is knowledge of the prototypicality status of objects, while in Experiments 2 and 3, the world knowledge being prioritized is information about the real-world size of objects.

3.2 Experiment 1a: Replication and method validation

Initial evidence for integration of size knowledge and information from the immediate visual context comes from the first experiment reported in Tribushinina (2011). This experiment was conducted in Dutch, but the findings should, in principle, generalize to other languages with similar adjective semantics, such as English. Tribushinina also conducted her experiments in the lab, which allows for tight control over the exact set-up the participants experience. The experimenter can determine the screen size, display resolution, and the distance between the screen and the participant. In an experiment relying on perception of size, these are factors that may potentially influence the results. My experiments are conducted in English and using MTurk, where I have no control over these factors relating to the participants' set-up. Therefore, before building on the paradigm used in Tribushinina (2011), I conducted a replication of her experiment (see Section 1.4.2), to validate these methodological changes.

Each trial displays a series of images, e.g., elephants, and the participant is asked "Which elephants are [big/small]?" Following Tribushinina (2011), there are two main predictions. The first prediction is that the mean number of images selected as *big* (the *big* zone) will be smaller than the mean number of images selected as *small* (the *small* zone). All seven of the images might be considered small if real-world size was the only information used to determine the comparison class, since all of the images are smaller than the objects they depict. If only information from the visual context were used, symmetry between *big* and *small* would predict an equal number of images considered *big* as considered *small*. It is not necessary for my argument to assume symmetry between *big* and *small*, though Tribushinina does take this stance.[2]

[2]It is plausible to assume that the function applied to the comparison class in order to determine the standard of comparison is the same for both *big* and *small*. Thus, if the visual display were the entire comparison class, then the result of applying the function would be symmetry between *big* and *small*, so that the *big* zone and the *small* zone are the same size. Even without assuming symmetry though, if world knowledge is not taken into account, the *big* zone should be the same across all objects, as should the *small* zone, since the images are the same size across all objects. This is shown not to be the case, both in Experiments 1 and 2, and even more so when the results

I predict that since speakers use both information from the immediate visual context and world knowledge, there will be an asymmetry between the two adjectives. The *big* zone will be **smaller** than the *small* zone because world knowledge informs the speaker that none of the images are actually big compared to other exemplars of this category previously experienced.

The second prediction relates to the prototypicality status of the objects. The depicted objects differed as to their prototypicality status, i.e. whether they were considered exemplars of prototypically big objects (e.g., elephants), prototypically small objects (e.g., babies) or neither exemplars of prototypically big nor prototypically small objects (designated here as neutral, e.g., cakes). Nevertheless, across objects, the images were the same sizes. Therefore, any differences in the number of images counted as *big* or as *small* between objects can be attributed to the influence of knowledge about their size, rather than to size information from the visual context. Thus, prototypically big objects are predicted to have a smaller *big* zone and a bigger *small* zone than prototypically small objects. Similar reasoning is used by Dirven and Taylor (1988) for why the collocation *tall skyscraper* is not rated as very acceptable, despite skyscrapers being prototypically tall (see Section 3.1). Neutral objects are predicted to fall in between the other two categories with regard to the size of their *big* and *small* zones.

3.2.1 Methods

Participants

A total of 57 people (36 male, 21 female) were recruited using Amazon's Mechanical Turk (MTurk). Self-selection was open to all and participants who were not native speakers of American English were excluded, as discussed in Section 3.2.2.

---

of these experiments are compared to those of Experiment 4, in Chapter 4. The differences I find in the zone size across objects demonstrate the integration of world knowledge with information from the immediate visual context in the determination of the comparison class. Thus, the assumption of symmetry between *big* and *small* is not necessary for the argumentation presented in interpreting the results of these experiments.

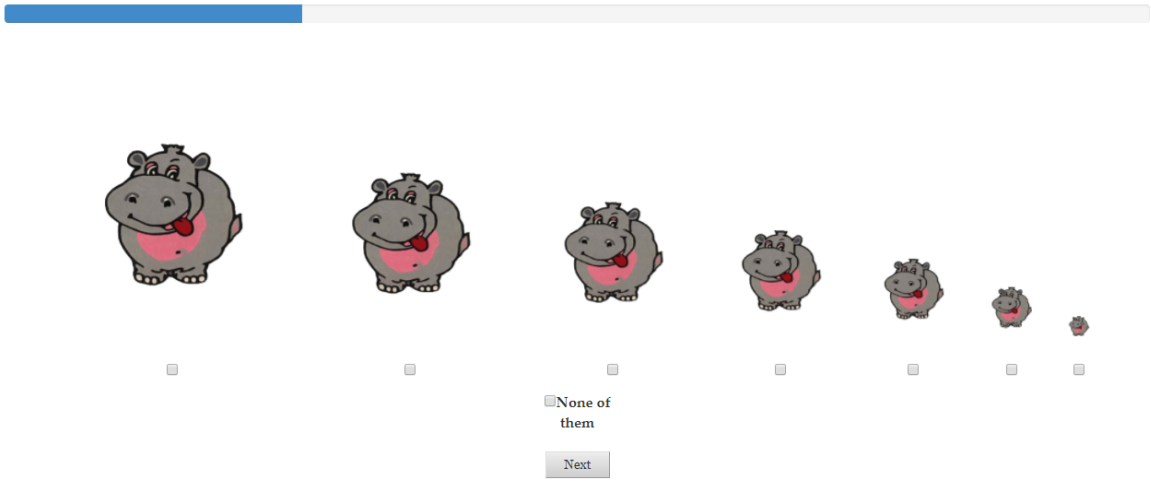


Figure 3.1: Sample trial display for Experiment 1a, showing hippos in ascending order. The display was accompanied by the audio question "Which hippos are big?"

Materials and Design

Each trial of the experiment consisted of a display containing seven clipart-style images of the same object, each scaled to a different size. The images were arranged by size, in either ascending or descending order. They ranged from 38 to 265 pixels in width, increasing by 38 pixels from image to image (Figure 3.1). Since the experiment was conducted using MTurk, the actual size of the displays that each participant saw varied between participants, as the images in the display scaled to fit the browser window. While the images scaled to fit the window, they remained consistent relative to each other, so that the relative proportions between each image and its neighbors were retained. The images used were identical to those used in Tribushinina (2011) and acquired from the author.[3] Twelve different images were used, four of prototypically big objects, four of prototypically small objects, and four of neutral objects, as specified in Table 3.1.[4]

[3]I am very grateful to Elena Tribushinina for sharing her original stimuli with me.

[4]For details on how the prototypicality status was determined, see Tribushinina (2011) and Tribushinina (2013a), summarized in Section 2.2.1. The norming study presented in Chapter 2 addresses the question of whether there are differences between English and Dutch in this respect. It provides evidence that there are differences in perceived size of objects of different prototypicality

Small	Neutral	Big
mouse	balloon	elephant
gnome	umbrella	plane
chick	cake	hippo
baby	monkey	house

Table 3.1: Prototypicality status of the objects displayed in Experiment 1a

As each set of stimulus images was displayed, the participant heard a question such as the example in (9), where the object matched the image on the screen. The audio stimuli were created using an online text-to-speech generator[5] in order to maintain consistent prosody across all of the trials. An image was selected either by clicking on the image or on the checkbox underneath it.

(9) a. Which [object]s are [big/small]?

b. Which [object]s do you find [big/small]?

c. Which [object]s seem [big/small] to you?

Procedure

The experiment consisted of two practice trials, followed by 48 randomized target trials, and finally, a brief demographic questionnaire. The two practice trials used objects (flower, car) and adjectives (*pretty*, *ugly*) not in the target trials. The purpose of these trials was to introduce participants to the task they were to perform and make them aware that they could select multiple answers. In both the practice phase and in the target trials, as each display appeared, the participant heard the audio prompt, and could not select any images until it was finished playing. The participant was required to select either at least one of the images or the "None of the above" box, before clicking on the "Next" button to advance to the next trial. At the end of the experiment, participants were asked for their native language and the size of the screen they used for the experiment along with other demographic information.

statuses in English, suggesting comparability to Dutch in this respect.

[5]https://text-to-speech-demo.mybluemix.net/

Each participant saw each object, in each order (ascending or descending), for each adjective (*big* or *small*), for a total of 48 target trials. The trials appeared in a randomized order with the constraint that the same object not appear twice in a row. Each participant only saw one verb, i.e. they saw sentences in only one of the templates found in (9).

Tribushinina (2011) and Tribushinina (2013a) used the phrasing *Welke olifanten vind je klein?*, corresponding to the English template in (9b). She chose this frame for two reasons. The first is that the verb *vinden* 'find' is the most common way to express subjective judgments in Dutch. The second reason is that in pilot studies using a template more like (9a) some participants thought the intention was to give consistent judgment across trials, whereas in normal language use production and interpretation of dimensional adjectives is subjective and specific to the situation. Therefore, in order to emphasize that subjective judgments were appropriate, the template equivalent to (9b) was used. In order to investigate whether this argument holds in English, Experiment 1a has three conditions, where the verb template is a between-subjects factor.

Exclusions

A number of criteria were used to ensure that the participants were attending to the task. All participants who reported using a screen smaller than 12 inches were excluded from the analysis, as were those who reported a native language other than English. Due to the scale structure of *big* and *small*, when asked to either select the big objects or the small objects, participants were expected to include the appropriate extreme image (either the biggest or the smallest). Participants who failed to select an appropriate endpoint more than 10% of the time were excluded. Also excluded were participants who provided non-consecutive responses[6] more than 10% of the time. Finally, all individual trials in which no appropriate endpoint was selected or in which non-consecutive images were selected were excluded.

[6]A response was non-consecutive, for example, if for *big* the largest and third largest images were selected but the second largest was not.

3.2.2 Results

In Experiment 1a, two participants were excluded for screen size, two for language and three for not selecting an endpoint on more than 10% of trials. Additionally, 26 individual trials were excluded.[7] After these exclusions, the data set contained 2,373 trials from 50 participants.

The data were analyzed using ordinal logistic regression models with planned orthogonal contrasts, calculated using the ordinal package (v. 2015.6-28) (Christensen, 2015) in R (v. 3.3.3) (R Core Team, 2017). These models are similar to performing a series of logistic regressions, respecting the ordered nature of the data – the sequential nature of the images and the monotonic size scale. The dependent variable was the number of images selected in each trial. This number is equivalent to the zone size of the adjective heard in that trial. The predictors included were the adjective mentioned in the trial, the prototypicality status of the object in the trial (prototypically big, small or neutral), the direction of the images in the display (ascending or descending), and the trial number, in order to see whether the results depended on where in the course of the experiment the trial appeared. Additionally, random effects were included for subject and item (object depicted, e.g., elephant).

The maximal model justified by the design was used where possible (Barr et al., 2013), and when it failed to converge, the structure of the subject random effects was reduced until convergence was achieved. An example of a model used appears in (10). The contrasts were set up to most directly compare the variables relevant to the hypotheses. First, the overall mean size of the *big* zone was compared to the overall mean size of the *small* zone. Next, within each zone, the zone size for prototypically big objects was compared to the combined mean of the neutral and prototypically small objects. Also within each zone, the zone size of the prototypically small objects was compared to the zone size of the neutral objects. The prototypicality status contrasts were designed to compare prototypically small objects directly to neutral ones since the real-world size of objects in these two categories is not necessarily different. Therefore, this is an important contrast for examining the effect of prototypicality

[7] Twelve trials were excluded for having no endpoint, four for having an inappropriate endpoint and eleven for having non-consecutive images selected.

knowledge on dimensional adjective interpretation.

(10) clmm(number_selected ~ adj.prototype_status + dir + question_order + (1+adj|noun) + (1+adj*prototype_status|subject_id), data=d.rep.are)

In order to examine verb choice an ordinal regression model was constructed with verb as an additional fixed effect. No model converged with any slopes in the random effect term for variation across subjects, so the term was reduced to contain only an intercept. Verb choice was not a significant predictor of number of images selected (Figure 3.2).

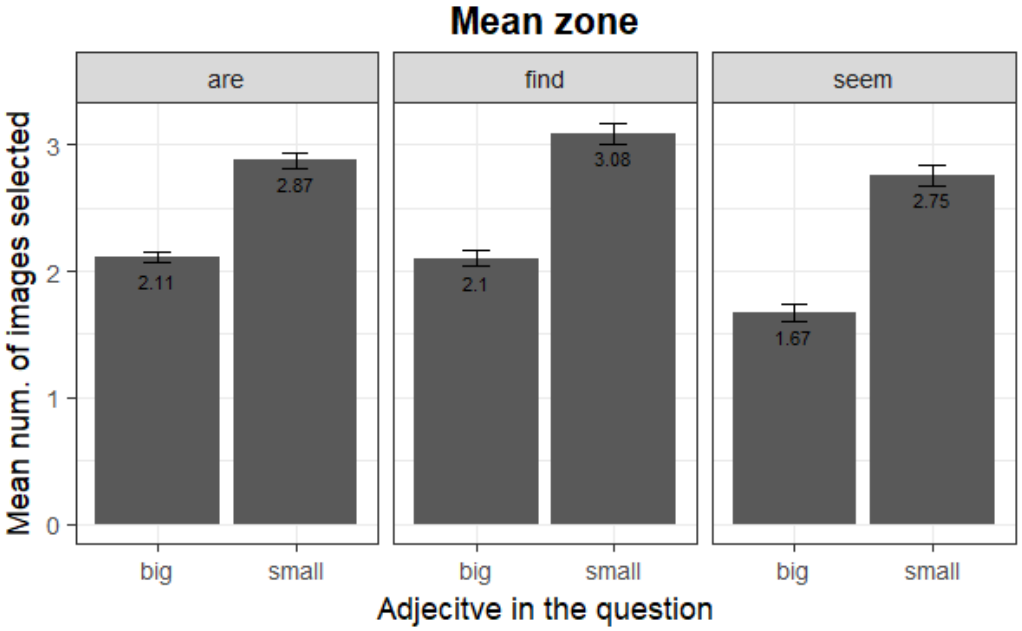


Figure 3.2: Mean number of images selected as *big* and *small* for each verb in Experiment 1a. Error bars represent 95% confidence intervals (CIs).

Comparing the template in (9a) using *are*, to the template in (9b) using *find*, there was no significant main effect of verb ($\beta = 0.25, p < 0.650$). Similarly, there was only a marginally significant difference between the template in (9a) using *are*, and the template in (9c), using *seem* ($\beta = -0.99, p < 0.065$). Looking at the differences in Figure 3.2, if anything, the difference between the mean number of images counted as

big (the *big* zone) and the mean number of images counted as *small* (the *small* zone) is smaller with *are*, and therefore harder to detect. Furthermore, semantically, *find* and *seem* introduce an element of subjectivity, and it is not obvious how that might interact with the adjective. *Are* is less semantically complex and therefore introduces fewer additional factors to contend with. Therefore, from this point forward, all further experiments are conducted using only the template in (9a). All further results for Experiment 1a will also be reported only for the condition using *are*, as in (9a).

Despite the methodological differences, I successfully replicated the results from Tribushinina (2011), where the *big* zone was found to be smaller than the *small* zone and protoypicality effects were observed. As predicted, I found that the *big* zone (mean = 2.11, sd = 0.81) is indeed smaller than the *small* zone (mean = 2.87, sd = 1.09). A Wilcoxon signed-rank test shows that this difference is significant ($Z = -21.44, p < 0.001$). Figure 3.3 shows the *big* and *small* zones for objects of each prototypicality status.

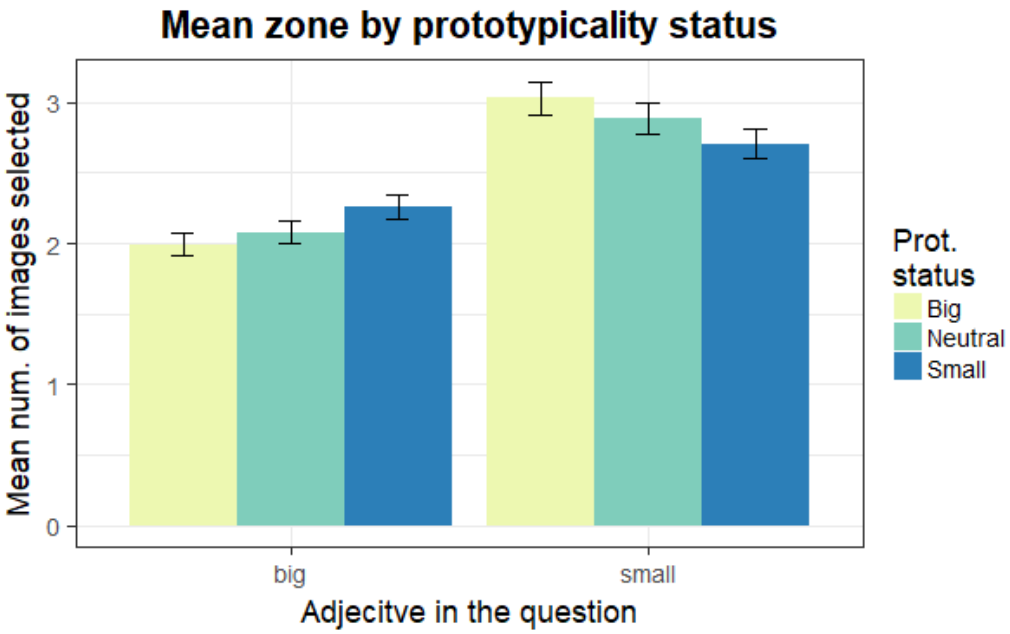


Figure 3.3: Mean number of images selected as *big* and *small* by prototypicality status of the object in Experiment 1a. Error bars represent 95% CIs.

The ordinal regression model for the *are* condition used an interaction between adjective and prototypicality status of the object as a random effect by subject.[8] Adjective has a significant effect on the number of images selected ($\beta = 4.14, p < 0.001$), such that when the adjective is *small* rather than *big*, significantly more images are selected. In other words, the *small* zone is significantly bigger than the *big* zone, as predicted in Tribushinina (2011).

Additionally, the data show evidence of prototypicality effects, as predicted. When the adjective in the question is *big*, the mean number of images selected for prototypically big objects (e.g., elephants) is significantly lower than the mean number of images selected for the prototypically small (e.g., babies) and neutral objects (e.g., cakes) combined ($\beta = -1.03, p = 0.025$). This suggests that the *big* zone is smaller for prototypically big objects than for other objects. Additionally, significantly fewer images are selected for neutral objects than for objects that are prototypically small ($\beta = -1.11, p = 0.042$).

A similar but inverse pattern is evident when the adjective in the question is *small*. The mean number of images selected is significantly higher for prototypically big objects than the mean number of images selected for the prototypically small and neutral objects combined ($\beta = 1.15, p = 0.004$). This result suggests that the *big* zone is smaller for prototypically big objects than for other objects. Additionally, very nearly significantly more images are selected for objects that are neither prototypically big nor small than for objects that are prototypically small ($\beta = 0.86, p = 0.052$). There is no effect of the direction the images were displayed in (ascending or descending) ($\beta = -0.116, p = 0.911$). There is a significant effect of question number (i.e. how far along in the experiment a question appeared) ($\beta = -0.029, p < 0.001$), such that the farther along in the experiment, the fewer images are selected, but it is a very small effect.

Broken down by object and arranged in order of number of images selected as *big/small* (i.e. zone size), the objects are mostly grouped by prototypicality status, as shown in Figure 3.4. For the most part, the prototypically small objects are grouped

[8]The model did not converge using the factor in the model which combines both adjective and prototypicality status into a single component, so a new model was constructed using two separate factors.

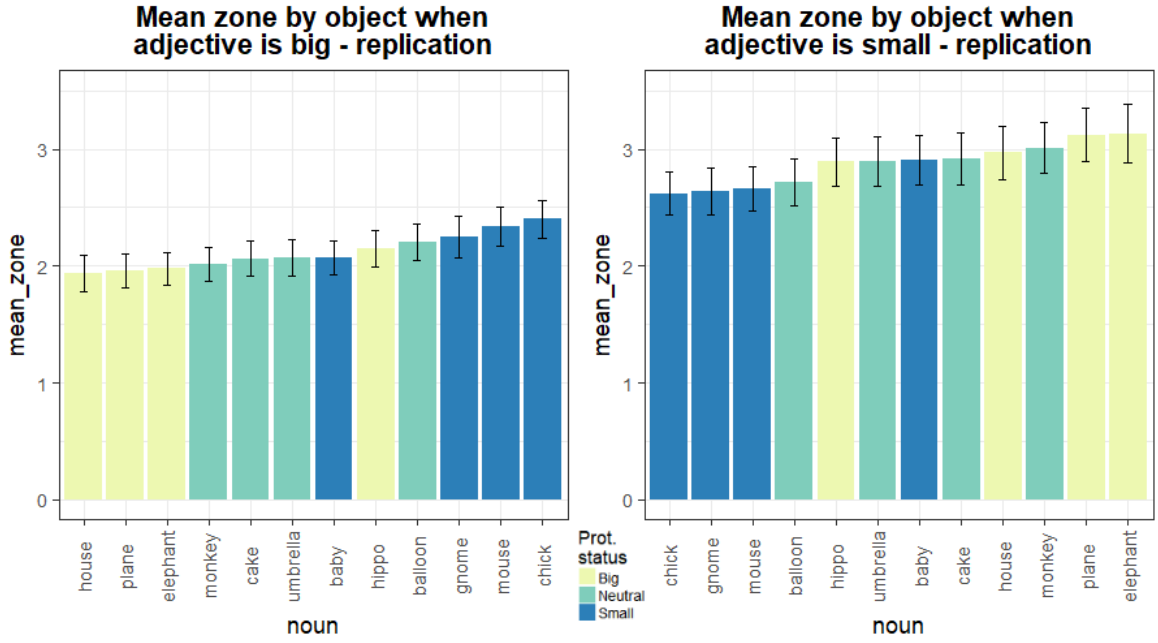


Figure 3.4: Mean number of images selected as *big* (left) and *small* (right) by object in Experiment 1a. Error bars represent 95% CIs.

at one end of each graph, and the prototypically big objects at the other. There are a couple of exceptions to this. With both adjectives, *baby* is separated from the other prototypically small objects and *hippo* is separated from the other prototypically big objects.

3.2.3 Discussion

Experiment 1a replicated the first experiment in Tribushinina (2011), but was conducted in English rather than Dutch and on MTurk rather than in the lab. Additionally, the effect of different verbs was examined. No significant difference was found between the condition with an objective verb like *be* in the question and the conditions with subjective verbs like *find* or *seem*. As the latter two introduce additional complications regarding the interaction of subjectivity and the interpretation of the adjective, in all of the following experiments, *be* is used as the verb in the stimulus questions.

The shift to MTurk required participants to perform the task on their own devices, which resulted in a lack of consistency of screen size, and thus stimulus size, across participants. I did exclude from analysis the participants who did the experiment on a mobile phone, since those images were significantly smaller than in the original experiment. Since they were being asked to judge size, the variation in the size of the stimuli across participants was a concern. Additionally, in the lab, there was an experimenter asking the questions, and the participants responded by pointing to the images they wanted to select. This interactional format might have led to more attentiveness or cooperativeness than on MTurk where participants may be multi-tasking and paying less attention to the task. On MTurk there was also the possibility of misclicking, or clicking without successfully selecting an image. These factors necessitated the exclusions described in Section 3.2.1, which were not required in the original experiment in Tribushinina (2011). Finally, semantic differences between English and Dutch might have affected the results.

Despite all these methodological differences, the results of Experiment 1a were similar to the findings in the first experiment in Tribushinina (2011). The results of Experiment 1a support the hypothesis that real-world size information is integrated with the information from the immediate visual context in determining the comparison class used for relative adjective interpretation. If the comparison class had been solely determined by the size of the images in the immediate visual context, then the *big* zone would have been expected to be the same size as the *small* zone. However, the *big* zone is significantly smaller than the *small* zone.

Additionally, across objects, each zone should be the same size, since each set of images were of the same sizes. In other words, prototypicality effects would not be expected to emerge. Assuming that the formula for determining the standard of comparison is fixed regardless of the distribution of items in the comparison class (e.g., Solt and Gotzner, 2012), then different sized zones indicate different distributions of sizes in the comparison class. Therefore, the observed difference between the size of the *big* zone for prototypically small objects and for prototypically big objects supports the hypothesis that other world knowledge is being used to determine the comparison class. However, the reasoning about this paradigm relies on all of the

images being smaller than the objects depicted in them. Not all of the objects strictly satisfy this criterion, a problem which is addressed in Experiment 1b.

3.3 Experiment 1b: Bunnies and ducklings

One of the premises of Experiment 1a is that all of the images are smaller than the objects depicted in them. The hypotheses and interpretation of the results are dependent on this premise. The *small* zone is expected to be bigger than the *big* zone if world knowledge is integrated with information from the immediate visual context because world knowledge determines that all of the objects in the displays are actually small relative to their size in the real world. However, the stimuli used do not all meet this requirement. The biggest images of the smallest of the prototypically small objects, mice and chicks, are actually approximately the size of actual mice and chicks.

In order to ensure that this deviation from the premise is not influencing the results, in Experiment 1b, the mice and chicks are replaced with bunnies and ducklings which, while still prototypically small, are larger in real life than mice and chicks. Therefore, even the biggest of the images depicting these objects used in this experiment is smaller than the object depicted.

If prototypicality status is being used in adjective interpretation, the two predictions remain the same as in Experiment 1a. First, the *small* zone is predicted to be larger than the *big* zone. Second, there should be prototypicality effects; the *small* zone is predicted to be bigger for prototypically big objects than for prototypically small ones, and the *big* zone is predicted to be smaller for prototypically big objects than for prototypically small ones. Objects that are neither prototypically big nor small are predicted to fall in between those in the other categories in both zones.

3.3.1 Methods

Participants

A total of 55 people (40 male, 15 female) were recruited using MTurk. Self-selection was open to all and participants who were not native speakers of American English were excluded, as discussed in Section 3.3.2. None of the participants took part in any of my previous experiments.

Materials and Design

The images used in Experiment 1b were identical to those from Experiment 1a, with two exceptions. The images for the mice and chicks were replaced with images of bunnies and ducklings as seen in Figure 3.5. The images were displayed in the same

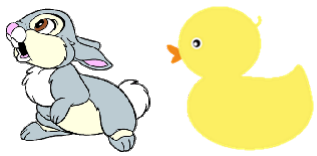


Figure 3.5: The bunny and duckling images used in Experiment 1b in place of the mouse and chick images

sets of seven images of ascending/descending size as in Experiment 1a. Since in Experiment 1a the verb choice did not have a significant effect and the option in (9a) is the most straightforward semantically, this template is used for the rest of the experiments reported in this chapter and the next. The audio prompts in Experiment 1b were the same as those used in Experiment 1a, with the relevant object changes, generated using the same web-based text-to-speech generator. A sample prompt appears in (11).

(11) Which bunnies are small?

Procedure

The procedure was identical to that of Experiment 1a. Each participant saw two practice trials (identical to those in Experiment 1a), followed by 48 randomized target trials, and finally, a brief demographic questionnaire. The trials appeared in random order, with the constraint that the same object could not appear twice in a row. Participants saw each object with each adjective with the images arranged both ascending and descending in size.

3.3.2 Results

Eleven participants were excluded from the analysis based on the exclusion criteria described for Experiment 1a in Section 3.2.1; four participants were excluded for using screens too small, two for having a native language other than English, and five for exceeding the threshold for trials with an inappropriate endpoint. Additionally, 25 individual trials were excluded,[9] leaving a total dataset of 2,087 trials over 44 subjects.

As predicted, since the images were all smaller than the objects depicted in them, the *small* zone (mean = 3.03, sd = 1.05) is bigger than the *big* zone (mean = 2.208, sd = 0.75). A Wilcoxon signed rank test shows that the difference between the zones is statistically significant ($Z = -20.04, p < 0.001$). The results (Figure 3.6) were further analyzed using ordinal regression models with planned orthogonal contrasts, as described in Section 3.2.2.

An ordinal regression model was constructed using an interaction between adjective and the prototypicality status of the object in the random effects term by subject, since the model using the single variable combining these components failed to converge, as in Experiment 1a. Adjective is a significant predictor of the number of images selected ($\beta = 4.32, p < 0.001$), such that when the adjective was *small* rather than *big*, significantly more images were selected. In other words, the *small* zone is significantly bigger than the *big* zone, as predicted in Tribushinina (2011).

[9] 13 trials were excluded for having no endpoint selected, 4 for having the wrong endpoint and 8 for including non-consecutive images.

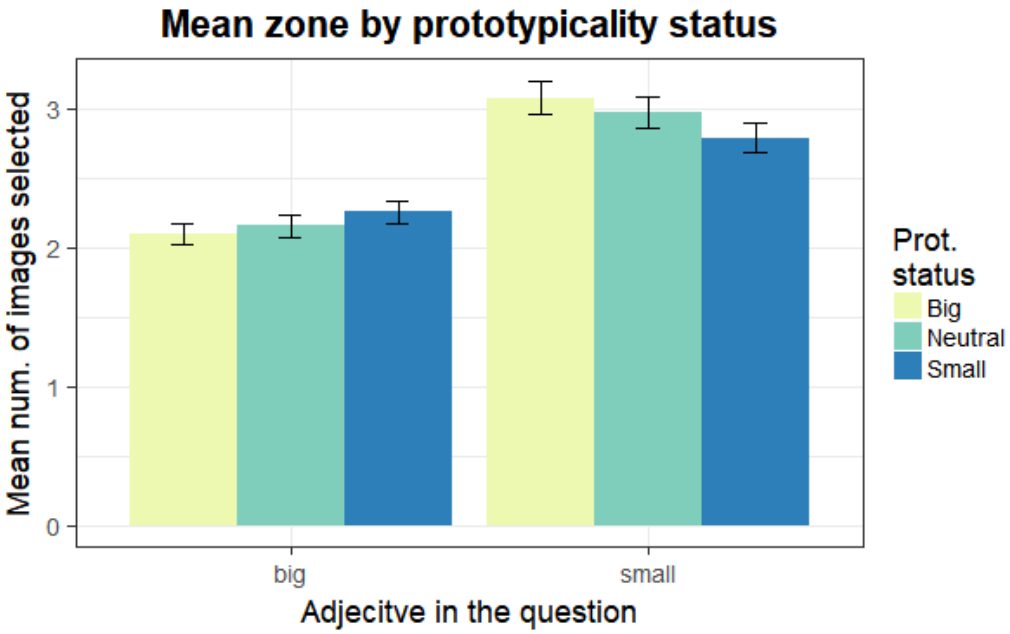


Figure 3.6: Mean number of images selected as *big* and *small* by prototypicality status of the object in Experiment 1b. Error bars represent 95% CIs.

Additionally, the data show some evidence of prototypicality effects, though not as clearly as in Experiment 1a. When the adjective in the question is *big*, prototypicality effects are not strongly evidenced. The mean number of images selected for prototypically big objects was only marginally significantly lower than the mean number of images selected for the prototypically small and neutral objects combined ($\beta = -0.66, p = 0.091$). This result suggests that the *big* zone tends to be smaller for prototypically big objects than for other objects. The number of images selected for neutral objects was not significantly different than the number of images selected for prototypically small objects ($\beta = -0.67, p = 0.152$), though the negative co-efficient shows that the effect is in the predicted direction.

However, when the adjective in the question is *small*, there is evidence of prototypicality effects. The mean number of images selected for prototypically big objects is significantly larger than the mean number of images selected for the prototypically small and neutral objects combined ($\beta = 1.06, p = 0.002$). This result suggests that

the *small* zone is bigger for prototypically big objects than for other objects. Additionally, significantly more images are selected for neutral objects than for objects that are prototypically small ($\beta = 1.05, p = 0.008$). There is no effect of the direction the images were displayed in (descending or ascending size) ($\beta = -0.07, p = 0.542$). There is also no effect of question number (i.e. how far along in the experiment a question appeared) ($\beta = 0.005, p = 0.232$).

Broken down by object, the objects are roughly grouped by prototypicality status, as shown in Figure 3.7. For the most part, the prototypically small objects are grouped at one end of each graph, and the prototypically big objects at the other, though there are more exceptions to this pattern than there were in Experiment 1a. With both adjectives, *monkey* is an outlier, behaving like the prototypically big objects, while *balloon* behaves like the prototypically small objects. When the adjective is *big*, *hippo* (which was an exception in Experiment 1a as well) is treated more like a prototypically small object than a big one. When the adjective is *small*, *hippo* behaves as predicted.

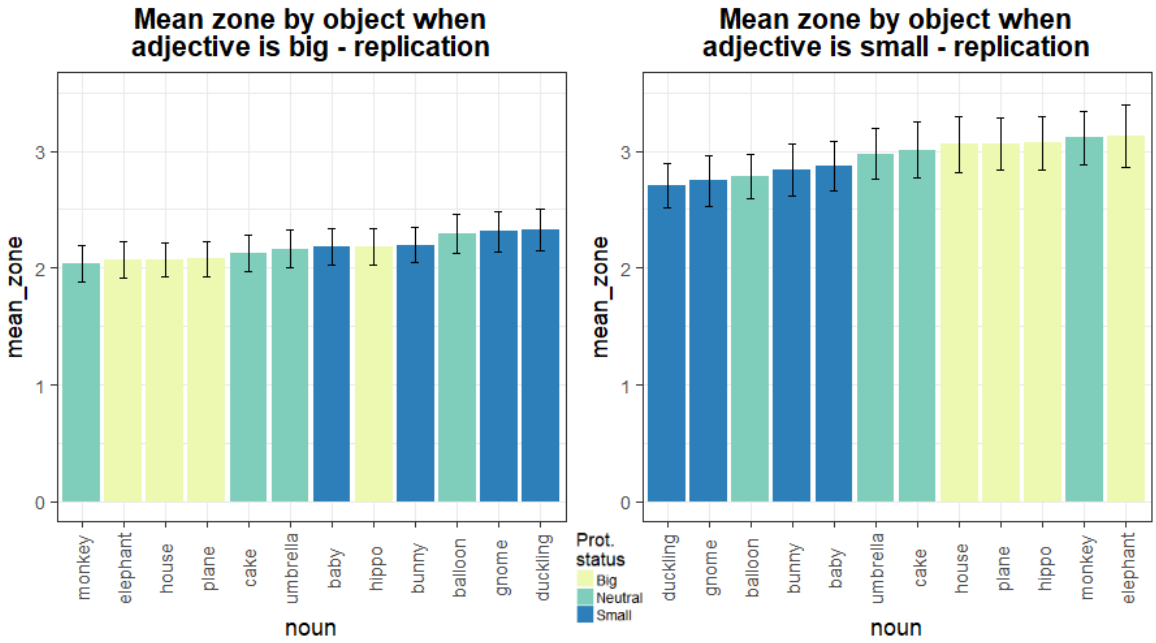


Figure 3.7: Mean number of images selected as *big* (left) and *small* (right) by object in Experiment 1b. Error bars represent 95% CIs.

3.3.3 Discussion

The results of Experiments 1a and 1b provide evidence that world knowledge is integrated with information from the visual context in order to determine the comparison class of a relative adjective such as *big* or *small*. While the prototypicality effects are not as clearly demonstrated as in Experiment 1a, there is still a significant difference between prototypically small and neutral objects when the adjective in the prompt is *small*.

The question remains, though, as to the nature of this world knowledge. It might be the case that a lexical item denoting a noun category such as *elephant* is associated with some fixed body of knowledge about entities in this category that is recalled every time the lexical item is mentioned. Figure 3.8 shows possible stimulus sets, the first, A, is a clipart depiction of elephants, as appeared in Experiments 1a and 1b. The second, B, is a photographic depiction.

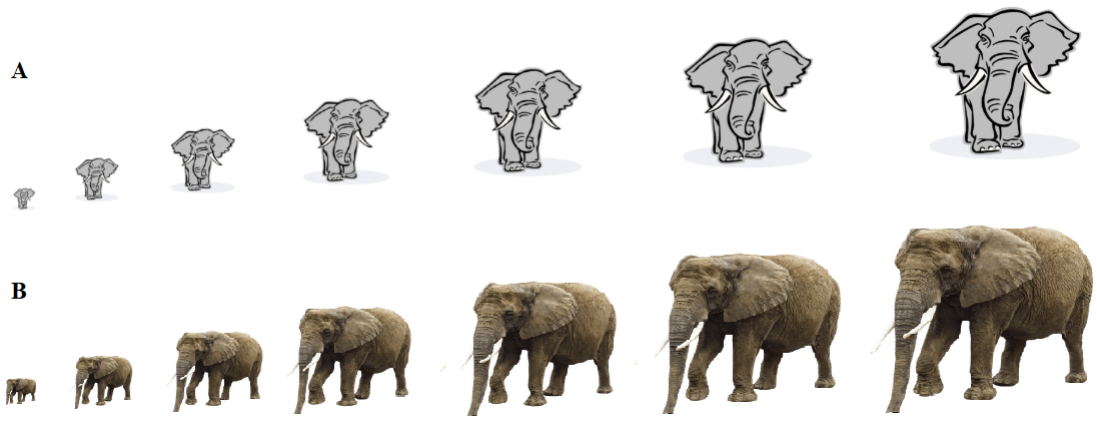


Figure 3.8: Two sample stimulus sets; Display A shows clipart stimuli depicting elephants while Display B shows photographic stimuli of elephants.

If the category *elephant* is associated with a fixed body of knowledge, then, given either set of elephants, regardless of the depiction type, the same sized elephants will count as *big*. For example, the two rightmost elephants might be selected in both displays. However, it is also possible that the type of elephant depiction matters. Different types of depictions of elephants (holding the size of the depiction constant)

might lead people to consider a different set of elephants as *big*. For example, in Display A, the two rightmost elephants might be selected as big, while in Display B, only the rightmost elephant might be selected. Assuming that the standard of comparison is calculated in a consistent manner, such a difference in the number of elephants counted as *big* suggests that there are different distributions of sizes of elephants in each comparison class.

In the next experiment, I examine the distinction between these two depiction types (and therefore context types): clipart-style depictions vs. photographic depictions. Clipart images do not necessarily depict an actual exemplar of the object category, while photographs, by their very nature, must depict an actual object that existed in the world. Clipart images may instead depict a generalized, idealized, typical object. Because of this difference, it may be that clipart images trigger access to a comparison class made up of representations of average, prototypical members of the category, which do not actually exist, and are not associated with actual dimensions. Rather, the representations in this comparison class may be generalized members of the class. Therefore, the distribution of sizes of these representations may be narrow and restricted.

Photographs, on the other hand, must be representations of actual objects which were originally associated with actual dimensions. They may trigger representations in the comparison class of actual objects of the depicted category which the speaker may have interacted with, and such objects would also have actual dimensions. These dimensions may be more widely distributed around the mean and contain more outliers.

Because of these differences, prototypicality status information may be more salient for clipart than for photographs. However, when judging photographs of objects, which more closely resemble the objects themselves than clipart renderings do, prototypical size information may be down-weighted as compared to real-world size information. The real-world size difference between the objects in the stimulus set for Experiments 1a and 1b that were categorized as prototypically small and the objects that were categorized as neutral is not very large. For example, a typical balloon is not very different in size from a typical baby, and in fact might be smaller than a baby.

Such examples show that there can be a misalignment between the prototypicality status and the actual size in the real world of the depicted object; in the real world a balloon might be smaller than a baby, despite babies being prototypically small and balloons not. Experiment 2 examines differences in context type, using a paradigm identical to the one used in Experiments 1a and 1b, except for photographs replacing the clipart images.

3.4 Experiment 2: Photographs

Clipart images emphasize stereotypical features, while photographs are depictions of objects as they exist in the world, associated with their real-world size. Differences between photographs and clipart have been previously investigated in the context of mapping language to visual referents. While this literature mostly focuses on questions of real-time language processing by examining the looking patterns at visual referents (e.g., Tanenhaus et al., 1995), there is some work on the effect the degree of realism has on the looking process (e.g., Saryazdi and Chambers, 2018, and references therein). The literature contains conflicting results as to whether more realistic pictures provide an advantage in processing.

In a task directly comparing carefully matched photograph and clipart stimuli, displaying four images on a screen where subjects needed to click on the object mentioned in a recorded sentence, Saryazdi and Chambers (2018) found that subjects' attention was held for a longer period of time after the stimuli appeared when the images were clipart rather than photographs. Additionally, when the noun was heard, subjects tended to identify the referent more quickly with clipart than with photographs. In a second experiment using a similar paradigm, participants encountered two types of verbs, "restrictive" verbs, e.g., *eat*, where two possible referents were edible and two were not, and "non-restrictive" verbs, e.g., *move*, which were semantically compatible with all of the objects depicted. No additional benefit for clipart in identification times was found when a specific verb helped disambiguate the referent.

Saryazdi and Chambers argue that, at least for mapping language to visual referents, there may be a small benefit to clipart over photographs, which is consistent

with "edge-based" accounts of object processing. These accounts hold that recognition of objects is driven mostly by the basic structural components of the object, and secondary features such as color and texture play at most a secondary role (Biederman, 1987). The simpler lines and clearer edges of clipart may be sufficient for object recognition while more information may only hinder the process.

In a norming study of 225 paired images consisting of a photograph and a clipart image generated directly from the photograph, Saryazdi et al. (2018) show that, across their norming measures, the two image types are, for the most part, comparable. One difference they found is in picture-name agreement, a measure of how well the label matched the image provided. The label was found to be a slightly better match to the object when the image was a photograph. They argue that this might be because the photo is a more faithful representation of the object depicted.

A difference was also found in a familiarity measure probing how much subjects interacted with or thought about the object, where photographs are rated significantly lower than clipart images. They speculate that photographs might be rated lower because clipart images typically are understood to depict a kind, whereas photos depict a specific instance of an object. While participants are likely to have interacted with the kind, they will not have interacted with the depicted instance. This difference may explain the lower ratings of photos, despite the fact that the question was phrased to specifically ask about interaction with the kind of object depicted.

Surprisingly, the clipart images are rated marginally more visually complex than the photographs. It is possible that this result reflects the complexity of the specific set of clipart images used in the experiment relative to other instances of clipart that participants might be familiar with. On measures of name agreement and variability,[10] associated verb agreement and variability,[11] and image agreement,[12] no differences

[10]Participants saw an image and were asked to provide the common English label for the object depicted. Scores were calculated to reflect agreement among participants on the name assigned, as well as the variability among the participants' answers.

[11]Participants were asked to provide the verb they judged to be most closely associated with the image. Scores were calculated to reflect agreement among participants on the most closely associated verb, as well as the variability among the participants' answers.

[12]Participants saw the name of an object, were asked to picture this object, then shown an image and needed to rate how well the image matched their mental picture.

were found between clipart and photographs. Overall, even on the measures in which they find a significant difference, the effect of image type is small.

In summary, previous literature has shown conflicting results on various measures of difference between clipart and photographs. The tasks used in these studies differ from the task in my experiments. Most previous studies used the image to aid processing, or examined image recognition. In my task the images form the context of the utterance and are crucial to its interpretation. I build on the different features of clipart and photographs noted by Saryazdi and Chambers (2018) and argue that, in this context, those differences give rise to different comparison classes when interpreting the adjective heard.

Clipart images emphasize stereotypical features, which I argue may lead a listener to place more emphasis on the prototypical size rather than the actual size of the object depicted when making size judgments. Therefore, I predict that the difference in the mean number of images selected as *big* or as *small* between prototypically small and neutral objects observed with clipart images should be attenuated or possibly disappear with photographs. Prototypically big objects are still predicted to be distinguishable from the other objects since the real-world size of the prototypically big objects in this study is many times the size of the other objects. There is no distinction in this case between real-world size and prototypical size.

Following this reasoning, the *big* zone is predicted to be smaller for prototypically big objects than for prototypically small or neutral objects, but the size of the *big* zone is not predicted to differ between prototypically small and neutral objects. Conversely, the *small* zone is predicted to be bigger for prototypically big objects than for prototypically small or neutral objects, but the size of the *small* zone is not predicted to significantly differ between the latter two categories. Since all images are still smaller than the objects they depict, the overall difference in size between the *big* zone and the *small* zone is not predicted to change from Experiment 1. Namely, the *big* zone is still predicted to be smaller than the *small* zone.

3.4.1 Methods

Participants

A total of 50 people (27 male, 23 female) were recruited using MTurk. Self-selection was open to all and participants who were not native speakers of American English were excluded, as discussed in Section 3.4.2. None of the participants took part in any of my previous experiments.

Materials and Design

Experiment 2 used photographs rather than clipart-style images, as shown in the sample display in Figure 3.9. The photographs show the same objects that appear in Experiment 1b (including the bunny and duckling introduced in Experiment 1b). The photographs are all close to square, so that when resized their aspect ratio is maintained, ensuring that the photographs of the different objects are actually the same sizes. This was not the case with the clipart.

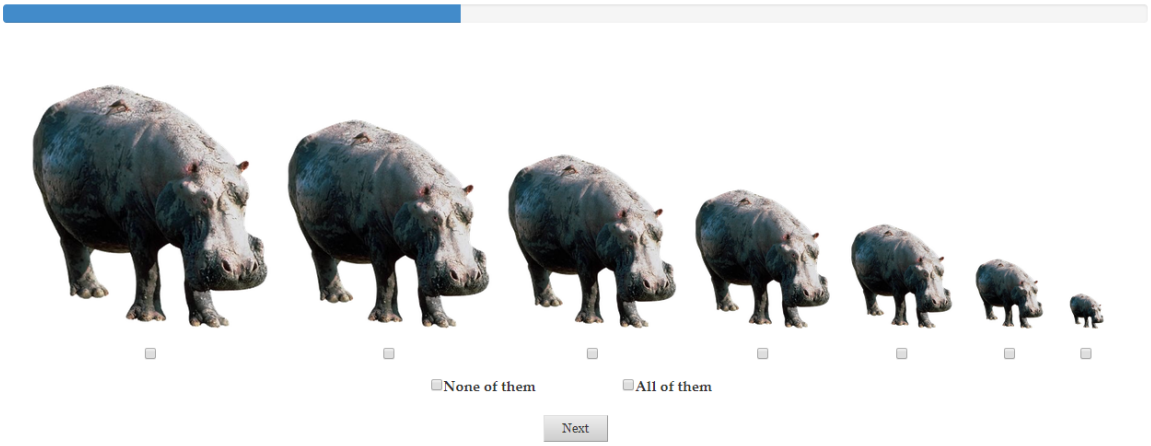


Figure 3.9: Sample trial display for Experiment 2, showing hippos in descending order. The display was accompanied by the audio question "Which hippos are small?"

Procedure

The procedure was identical to that of Experiments 1a and 1b. Each participant first saw two practice trials, which were also photographs, as opposed to the clipart seen in Experiments 1a and 1b. However, they used the same objects (cars, flowers) and adjectives (*pretty*, *ugly*) as in the previous experiments. These trials were followed by 48 target trials, in a random order, with the constraint that the same object could not appear twice in a row. Finally, participants were asked for demographic information.

3.4.2 Results

Nine participants were excluded from the analysis based on the exclusion criteria described for Experiment 1a in Section 3.2.1. Five participants were excluded for using screens that were too small, and four for exceeding the threshold for trials with an inappropriate endpoint. Additionally, 35 trials were excluded,[13] leaving a total dataset of 1,933 trials over 41 subjects.

As predicted, since the images are all smaller than the objects they depict, the *big* zone (mean = 2.33, sd = 0.8) is smaller than the *small* zone (mean = 3.14, sd = 0.8). A Wilcoxon signed rank test shows that the difference between the zones is statistically significant ($Z = -16.16, p < 0.001$). The results (Figure 3.10) were further analyzed using ordinal regression models with planned orthogonal contrasts, as described in Section 3.2.2.

An ordinal regression model with the maximal random effects structure justified by the experimental design was constructed. Adjective has a significant effect on the number of images selected ($\beta = 4.02, p < 0.001$), such that when the adjective is *small* rather than *big*, significantly more images are selected. In other words, the *small* zone is significantly bigger than the *big* zone, as predicted in Tribushinina (2011).

Interestingly, as opposed to the data patterns observed in Experiments 1a and 1b, these data do not provide evidence of prototypicality effects with photographic images as stimuli. When the adjective in the question is *big*, prototypicality effects are not

[13] 22 trials were excluded for having no endpoint selected, 3 for having the wrong endpoint and 10 for including non-consecutive images.

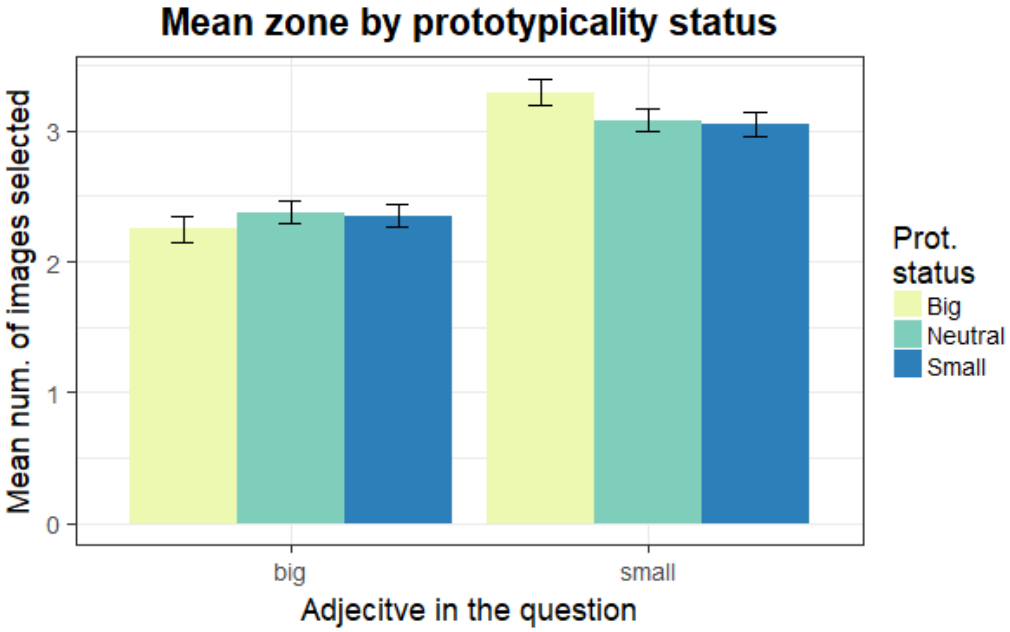


Figure 3.10: Mean number of images selected as *big* and *small* by prototypicality status of the object in Experiment 2. Error bars represent 95% CIs.

strongly evidenced. The mean number of images selected for prototypically big objects is only marginally significantly lower than the mean number of images selected for the prototypically small and neutral objects combined ($\beta = -0.63, p = 0.094$). This result suggests that the *big* zone tends to be smaller for prototypically big objects than for other objects. The number of images selected for neutral objects is not significantly different than the number of images selected for prototypically small objects ($\beta = -0.05, p = 0.896$).

When the adjective in the question is *small* there is no evidence of prototypicality effects. The mean number of images selected for prototypically big objects is significantly higher than the mean number of images selected for the prototypically small and neutral objects combined ($\beta = 1.09, p = 0.035$). This result suggests that the *small* zone is bigger for prototypically big objects than for other objects. As opposed to the results of Experiments 1a and 1b, there is no significant difference between

the number of images selected for neutral objects and for objects that are prototypically small ($\beta = 0.09, p = 0.827$). There is no effect of the direction the images are displayed in (descending or ascending) ($\beta = -0.02, p = 0.861$). There is also no effect of question number (i.e. how far along in the experiment a question appeared) ($\beta = 0.004, p = 0.268$).

Broken down by object, the objects of differing prototypicality statuses are more interleaved than in the previous experiments, as shown in Figure 3.11. For the most part, the prototypically big objects are separated from the other objects, while the neutral and prototypically small objects are fairly interleaved. With both adjectives, *baby* is an outlier, as it was in Experiment 1a, behaving like the prototypically big objects.

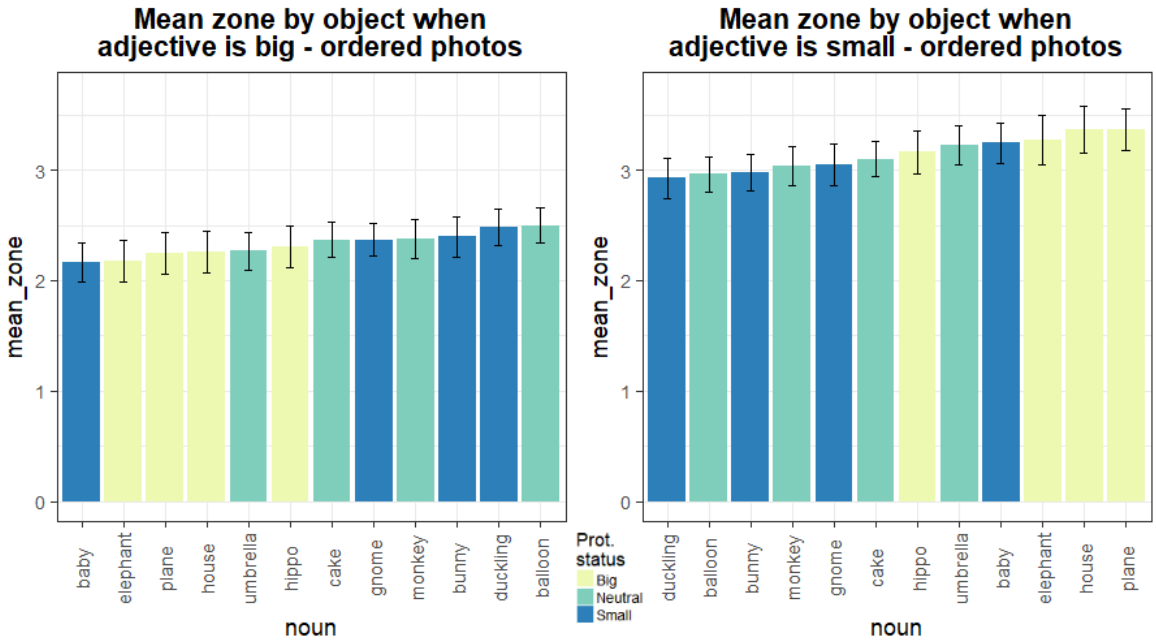


Figure 3.11: Mean number of images selected as *big* (left) and *small* (right) by object in Experiment 2. Error bars represent 95% CIs.

About ten percent of the subjects seem to have simply formulated a response strategy in which they choose, for example, the two biggest or the two smallest objects in each display in response to the adjective heard, as seen in Figure 3.12, where the

number of images selected in each trial is plotted for each participant. A flat line indicates that the same number of images were selected throughout the experiment regardless of the object and adjective.

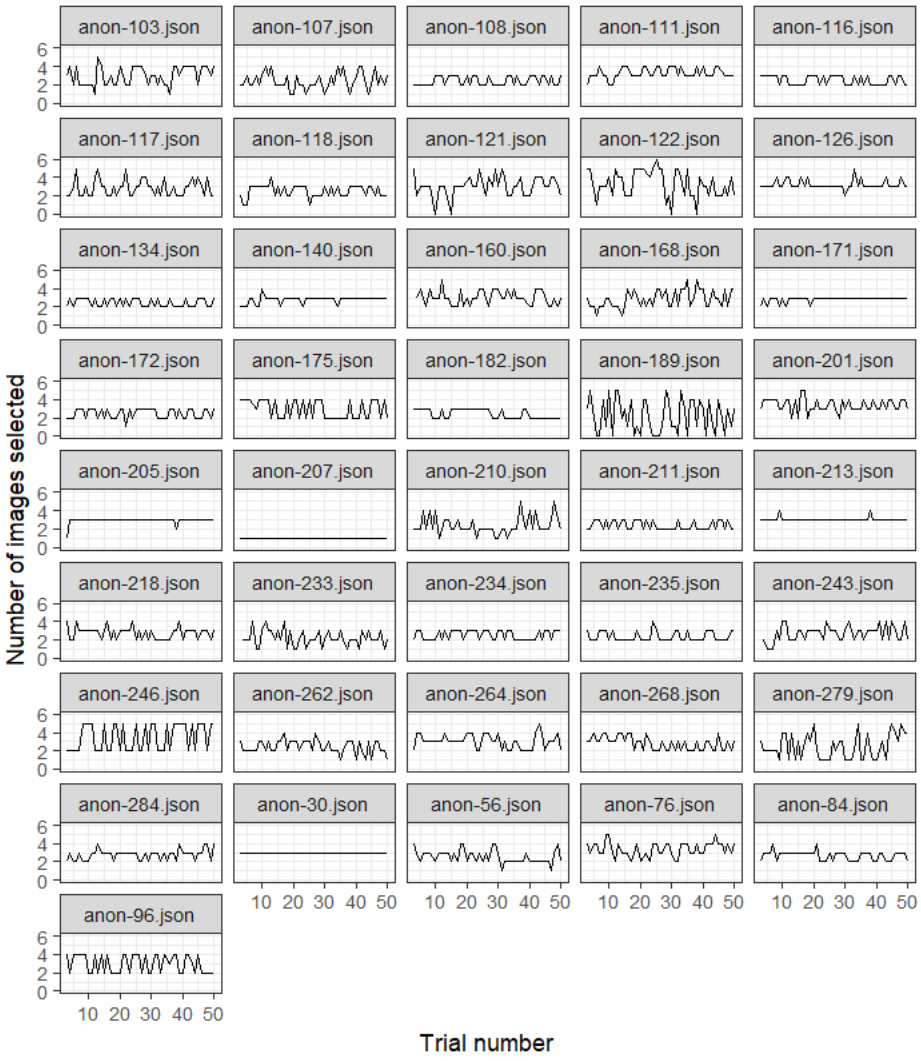


Figure 3.12: Number of images selected by each participant in each trial throughout Experiment 2

3.4.3 Discussion

The pattern of significant differences found in Tribushinina (2011) and in Experiments 1a and 1b is not replicated with photographic images. Experiment 2 does not provide evidence for prototypicality effects. No significant difference is found in zone size between the prototypically small and neutral objects. This result is consistent with an interpretation that when speakers judge the size of objects in photographs, prototypicality status is not taken into account, as opposed to when they judge clipart. When judging photographs, the difference between prototypically small and neutral objects is no longer significant. This difference between the results in Experiments 1a and 1b as compared to Experiment 2 may be driven by the lack of difference in the real-world size of the objects in these two categories, whereas there is a difference between their size and the size of the prototypically big objects.

The move from clipart to photographs was made in part to ensure that the results hold when the stimuli more closely resemble things that might be encountered in a naturalistic setting. The ordering of the stimuli in each display by order of size is also far from naturalistic, since in the real world the objects one interacts with are not typically ordered by size. Additionally, the ordering may encourage participants to simply formulate a response strategy in which they choose a fixed number of images to select, e.g., the two biggest/smallest, and adhere to it throughout the experiment.

In order to more closely approximate a naturalistic setting, as well as confound attempts to respond based on a fixed strategy instead of evaluating each display on its own merits, I conducted Experiment 3. This experiment uses the images from Experiment 2, but the order of the seven images in each display is randomized, instead of ordered by either ascending or descending size.

3.5 Experiment 3: Randomized order

In the paradigm used for Experiments 1a – 2 the images in each display are ordered by size. The ordering may influence participants' response strategies so that they do not examine each display on its own merits but rather choose a strategy for answering

the question and then automatically apply it. Randomized stimuli are also more naturalistic, as adjectives in natural speech are not often used in the context of neatly ordered objects. For example, faced with a group of zebras at the zoo, one might point out the big one. In this scenario, the zebras are unlikely to be standing neatly ordered by size, and yet the listener has no trouble picking out the relevant zebra.

Experiment 3 takes a more naturalistic perspective on the questions previously asked, randomizing the order of the stimuli in each display. I do not expect to find significant differences between the results of Experiments 2 and 3. I still predict that the *big* zone will be smaller than the *small* zone, since the images are all still smaller than the objects depicted in them. Additionally, as in Experiment 2, photographs of the objects were used; therefore, I do not expect to find a distinction between prototypically small and prototypically neutral objects. However, these two categories are still expected to be distinguishable from the prototypically big objects, since the latter are significantly bigger in actual size as well as being prototypically big.

3.5.1 Methods

Participants

A total of 103 people (60 male, 43 female) who had not participated in Experiments 1a, 1b or 2 were recruited using MTurk.[14] Self-selection was open to all and participants who were not native speakers of American English were excluded, as discussed in Section 3.5.2.

Materials and Design

Experiment 3 used the same photographic images as Experiment 2. Instead of the images being arranged in ascending or descending order (from left to right), for each trial the image order was randomized (Figure 3.13).

[14]Since the stimuli in this experiment appeared in a randomized order, there was no factor of direction of display and therefore, half the number of trials as in the previous experiments. Double the number of people as in the previous experiments were recruited in order to achieve the same power.

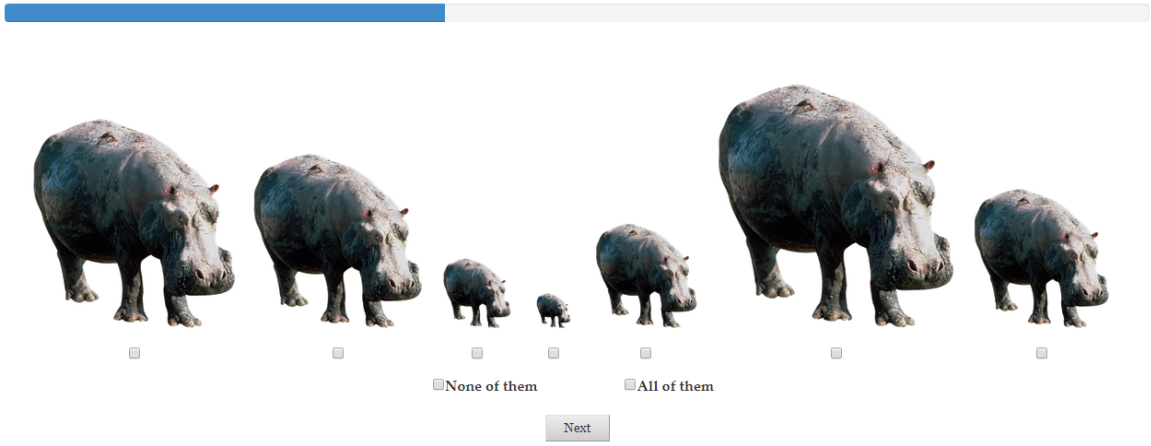


Figure 3.13: Sample trial display for Experiment 3, showing hippos in randomized order. The display was accompanied by the audio question "Which hippos are small?"

Procedure

The procedure is similar to that of Experiment 2. The experiment started with two practice trials identical to those in Experiment 2. The images in these trials were not in any particular order in either case, since the adjectives used are associated with a subjective scale (*pretty* and *ugly*). These practice trials were followed by 24 test trials, in a random order, with the constraint that a given object could not appear in consecutive trials. In Experiment 3, participants saw only 24 test trials (as opposed to 48), one for each adjective with each object. Finally, as in the previous experiments, participants were asked to provide demographic information.

3.5.2 Results

Eleven participants were excluded from the analysis based on the exclusion criteria described for Experiment 1a in Section 3.2.1; five participants were excluded for using screens that were too small, one for having a native language that is not English, three for exceeding the threshold for trials with an inappropriate endpoint, and two for exceeding the threshold for non-consecutive trials. Additionally, 59 trials were

excluded,[15] leaving a total dataset of 2,203 trials from 92 subjects. The participant exclusion criteria may be harsher than necessary given the change in the task from Experiments 1a, 1b and 2. With the images randomized it is easier to accidentally not select consecutive images, or to miss the endpoint and, therefore, even having ten percent of a given error type might not not indicate lack of cooperation, but rather lack of attention. Nevertheless, the criteria were kept the same in order to provide consistency with the previous two experiments.

As predicted, for images that are smaller than the objects they depict, the *big* zone (mean = 2.49, sd = 0.75) is smaller than the *small* zone (mean = 2.89, sd = 0.8). A Wilcoxon signed rank test shows that the difference between the zones is statistically significant ($Z = -12.94, p < 0.001$). The results (Figure 3.14) are further analyzed using ordinal regression models with planned orthogonal contrasts, as described in Section 3.2.2.

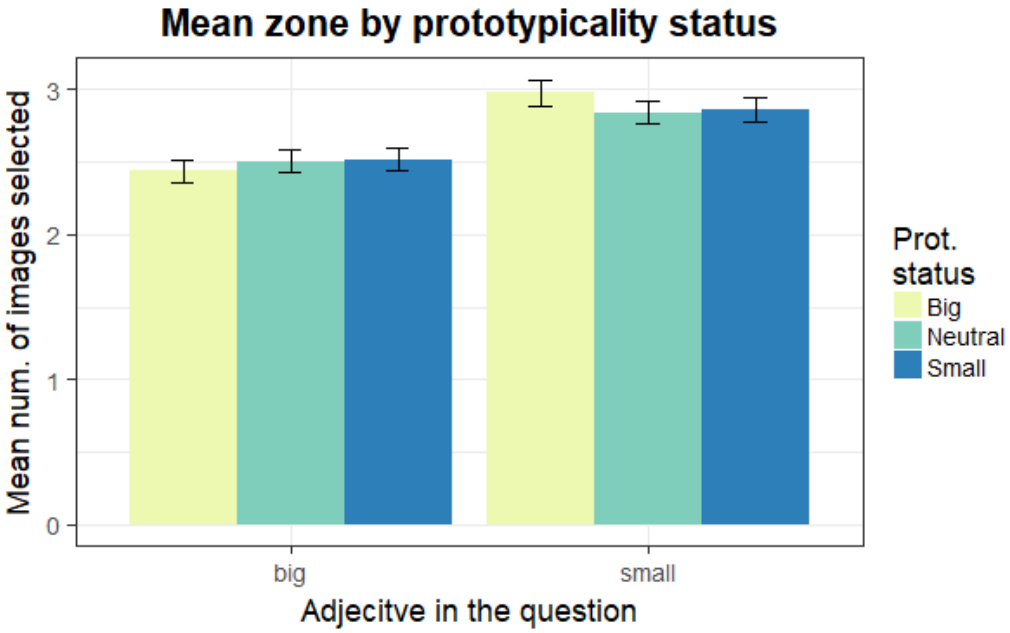


Figure 3.14: Mean number of images selected as *big* and *small* by prototypicality status of the object in Experiment 3. Error bars represent 95% CIs.

[15] 32 trials were excluded for having no endpoint selected, 2 for having the wrong endpoint and 25 for including non-consecutive images.

An ordinal regression model was constructed which only has slopes by adjective included in the random effect for subjects, due to non-convergence of the models with more complex random effects structures. Adjective has a significant effect on the number of images selected ($\beta = 1.74, p < 0.001$), such that when the adjective was *small* rather than *big*, significantly more images were selected. In other words, the *small* zone is significantly bigger than the *big* zone, as predicted.

In Experiments 1a and 1b, when the stimuli are clipart, prototypicality effects are observed. The *big* zone is smaller for prototypically big objects than for prototypically small ones, while the *small* zone is smaller for prototypically small objects than for prototypically big ones. In Experiment 2, when the stimuli are ordered photographs, these effects break down somewhat. The zone sizes are not significantly different for prototypically small objects than for neutral objects. This pattern is further strengthened by the results of Experiment 3.

When the adjective in the question is *big*, there is no evidence of prototypicality effects. The mean number of images selected for prototypically big objects is not significantly different than the mean number of images selected for the prototypically small and neutral objects combined ($\beta = -0.31, p = 0.292$). The difference in the number of images selected for neutral objects than for objects that are prototypically small is also not significant ($\beta = -0.04, p = 0.896$).

When the adjective in the question is *small* there is also no evidence of prototypicality effects. The mean number of images selected for prototypically big objects is not significantly different than the mean number of images selected for the prototypically small and neutral objects combined ($\beta = 0.54, p = 0.139$). As in Experiment 2, the difference between the number of images selected for neutral objects and for objects that are prototypically small is not significant ($\beta = -0.15, p = 0.710$). There is a small but significant effect of question number (i.e. how far along in the experiment a question appeared) ($\beta = 0.019, p = 0.008$).

Broken down by object, the objects of differing prototypicality statuses are fairly interleaved, as in Experiment 2, as shown in Figure 3.15. With both adjectives, *baby* is again an outlier, as in Experiments 1a and 2, behaving like the prototypically big objects.

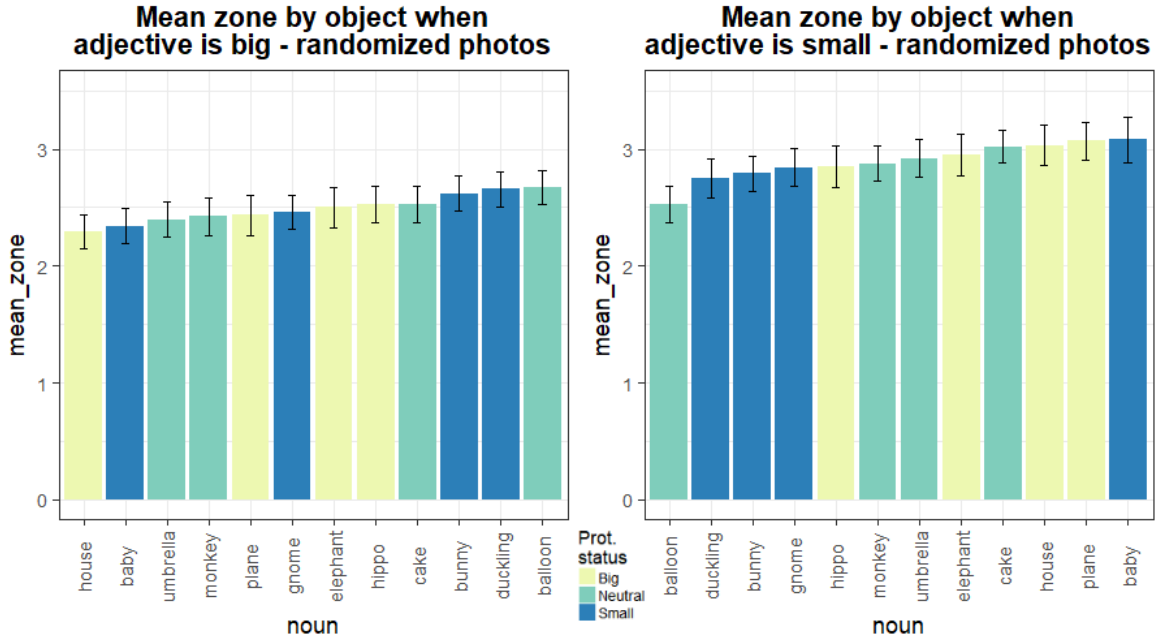


Figure 3.15: Mean number of images selected as *big* (left) and *small* (right) by object in Experiment 3. Error bars represent 95% CIs.

Figure 3.16 shows the number of images selected in each trial plotted for each participant, and a flat line indicates that they chose the same number of images regardless of the object and adjective. Despite my predication that randomizing the order of the images in the display would minimize the formation of response strategies, a non-negligible number of participants still seem to have selected a strategy for answering the questions, at least by midway through the experiment. Unlike in Experiment 2, there were no participants who selected the same number of images for every trial, but some participants did settle into a fixed response pattern fairly early on (e.g., anon-226 in Figure 3.16b).

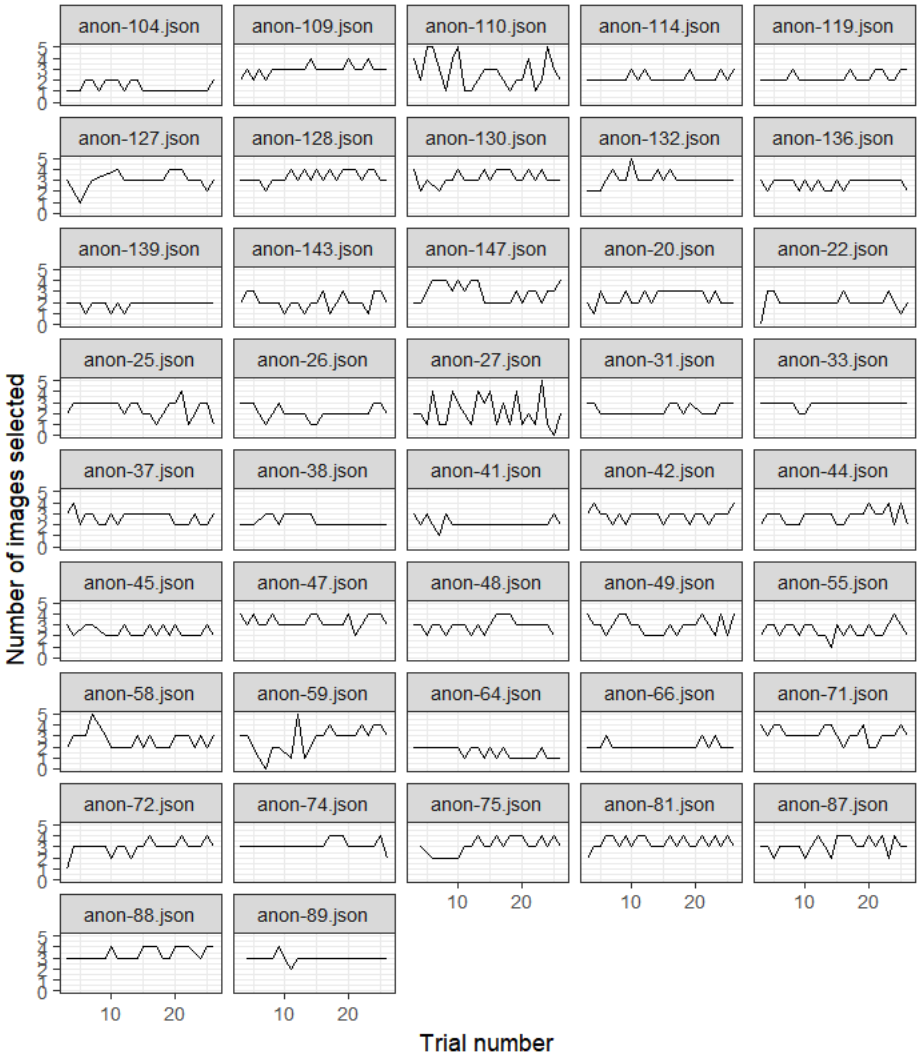


Figure 3.16a: Number of images selected by each participant in each trial throughout Experiment 3 for half of the participants

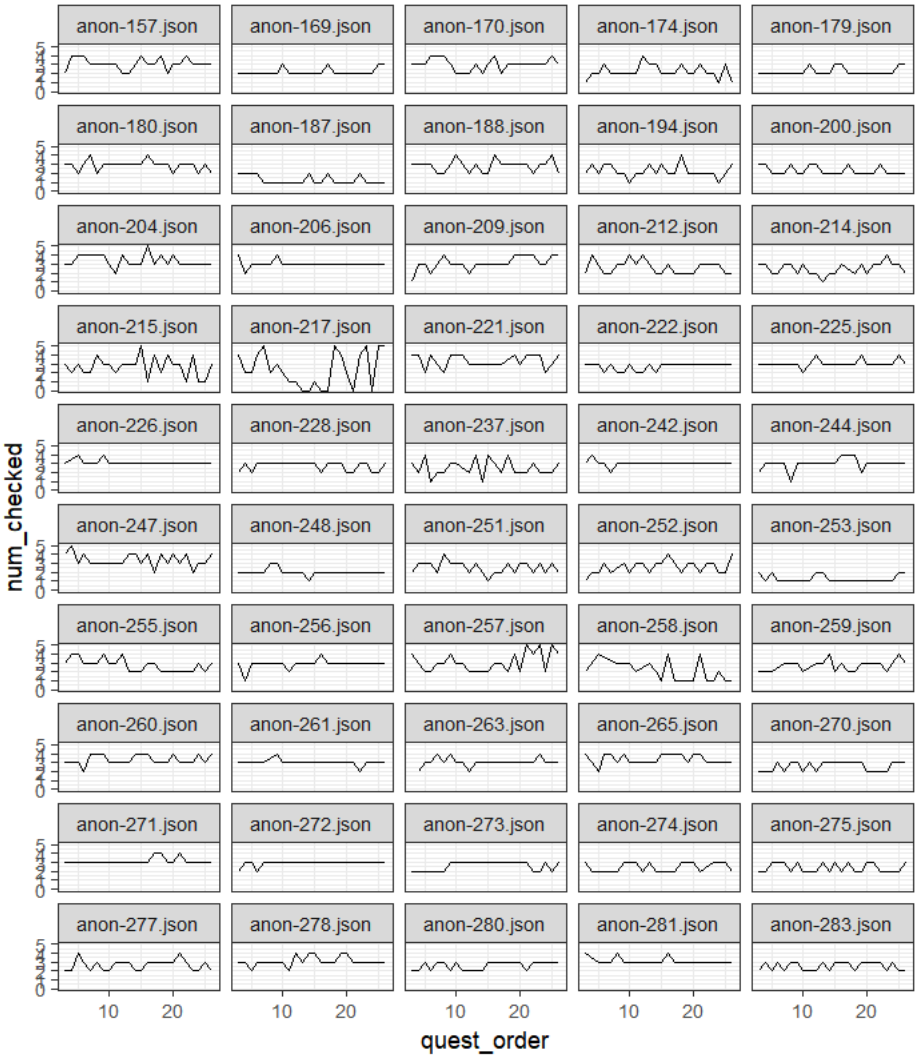


Figure 3.16b: Number of images selected by each participant in each trial throughout the experiment for Experiment 3 for the other half of the participants

3.5.3 Discussion

In Experiments 1a, 1b and 2 the stimuli seen by the participants were arranged in order of size. This set-up raised the concern that participants might simply be choosing a response strategy, such as to select the largest or smallest two images for every trial, without thinking about the question they were asked in relation to each

set of stimuli. The ordered stimuli are also not at all naturalistic; people do not often come across sets of things organized neatly by size. Therefore, Experiment 3 presented participants with randomized sets of images. Some of these orders ended up being fairly organized, and others less so, but over the course of the experiment, and across subjects, the effect of any specific ordered set of stimuli should be minimized. The introduction of randomized ordering should make it more difficult for a participant to select a simple strategy since they would then have to go to the added effort of reordering the stimuli in their heads in order to pick the top or bottom two. It would be less effortful to simply decide which were big or small for that set.

As is evident in Figure 3.16, some participants still seem to have selected a strategy for answering the questions, at least by midway through the experiment. It is possible that the initial reasoning should be reversed. It is possible that because of the additional cognitive complexity in re-ordering the images, participants are more likely to choose a fixed response strategy, reducing the cognitive burden of providing an answer.

The results of this experiment strengthen the pattern with regard to prototypicality effects observed in Experiment 2. There is no difference between the size of the *big* zone or the size of the *small* zone for prototypically small objects as compared to neutral objects. Surprisingly, the difference between the prototypically big objects and the other objects is not significant either in this experiment. This is unexpected if the real-world size is being taken into account more than prototype information for photograph stimuli, since the real-world size of the prototypically big objects is significantly larger than that of the other objects, as evidenced by the norming study reported on in Chapter 2. It is possible that the randomization introduced enough noise that the effect of real-world size is now not detected. It is also possible that the data should be filtered to exclude participants who chose a fixed or almost fixed response strategy, and that if these participants are removed an effect might emerge.

The main effect of zone size is maintained as predicted. The *big* zone is smaller than the *small* zone, which is consistent with the hypothesis that speakers integrate world knowledge of the size of these objects along with information from the immediate visual context when making adjectival judgments. This effect is smaller in

this experiment than in Experiments 1a, 1b and 2, perhaps due to the introduction of noise by randomizing the order of the stimuli. However, it is still statistically significant.

3.6 General Discussion

In this chapter, I presented a series of four experiments. Experiment 1a replicated the first experiment in Tribushinina (2011). I conducted this experiment in English rather than in Dutch, and on MTurk rather than in a lab. This study served as a cross-language, online replication, validating the methodology changes I made and laying the foundation for further experiments. I found that overall fewer images were selected as *big* than as *small*. The data also displayed prototypicality effects, so that the *big* zone was bigger for prototypically small objects than for prototypically big ones, and the *small* zone was bigger for prototypically big objects than for prototypically small ones. Both zone sizes for objects that were neither prototypically big nor prototypically small fell in between the other two prototypicality status groups. In Experiment 1b, two of the prototypically small objects were replaced to ensure that all of the stimulus images were smaller than the objects depicted. The results were similar to those of Experiment 1a; the *big* zone was smaller than the *small* zone, and there was evidence of prototypicality effects. In Experiment 2, the clipart stimuli from Experiment 1b were replaced by photographic images of the same objects. Across all objects, as in Experiments 1a and 1b, the *big* zone was smaller than the *small* zone. However, the difference in the size of each zone for prototypically small objects was not distinct from that of neutral objects. The results of Experiment 3 were similar to those of Experiment 2 for photographic stimuli displayed in random order (i.e. not ordered according to size).

Taken together, these studies provide evidence supporting the two hypotheses that were presented in Section 3.1. The first hypothesis I proposed is that world knowledge is integrated with information from the immediate visual context when determining the comparison class for relative adjective interpretation. The second hypothesis is that world knowledge is not a single monolithic body that exerts exactly the same

influence in every instance of adjectival judgments. With these experiments, I provide support for a distinction between two types of world knowledge: knowledge about an object category's prototypicality status, and knowledge about the real-world size of objects. Furthermore, I argue that contextual factors, such as the image type of the experimental stimuli, determine the interaction between these types of knowledge in forming the comparison class relative to which an adjective is interpreted.

In Experiments 1–3, the *big* zone was consistently smaller than the *small* zone, regardless of the object depicted. This result supports the first hypothesis: world knowledge is integrated with the information from the immediate visual context in determining the comparison class. The observed asymmetry suggests that when speakers interpret relative adjectives, they integrate world knowledge with the information from the immediate visual context in some way in order to form the comparison class relative to which an object is evaluated.

The reasoning behind this hypothesis is the following. Were the size information from the immediate context the only source of information being taken into account, the *big* zone and the *small* zone would be expected to be the same size.[16] However, speakers also take into account knowledge they have about the real-world size of the objects depicted in the images. They know that all of these images are smaller than the real-world size of the objects depicted in them. It is, therefore, harder to count as a *big* object, since it is known that the image is actually smaller than the object depicted in it. Moreover, if world knowledge regarding the size of the depicted images were the only source of information being taken into account, then none of the objects should count as *big*. The selection of any images at all as *big* indicates that the information from the immediate visual context is also taken into account when determining the comparison class. Therefore, there seems to be an integration of the size information of the images in the immediate visual context along with prior world knowledge regarding the objects depicted in order to form a comparison class to interpret the adjective heard.

[16]This argument only holds if there is no inherent asymmetry in the semantics of the two adjectives. Chapter 4 will argue against this alternative explanation.

I also found evidence to support the second hypothesis: two types of world knowledge can be distinguished, and contextual factors determine the interaction between them in affecting the comparison class. Prototypicality status did not have the same effect in Experiments 2 and 3 as it did in Experiments 1a and 1b. In both Experiment 1a and 1b, I found the same prototypicality effects as observed in Tribushinina (2011). In Experiment 1a, the *big* zone was smaller for prototypically big objects than for neutral objects and smaller for neutral objects than for prototypically small objects. Conversely, the *small* zone was bigger for prototypically big objects than for neutral objects, and bigger for neutral objects than for prototypically small objects. In Experiment 1b the effects were less evident, but the *small* zone was significantly smaller for prototypically small objects than for neutral objects. Elephants in general are considered *big* and babies in general are considered *small*, and therefore it is harder to count as a *big elephant* than it is to count as a *big baby*. A baby needs to be less big relative to other babies (in the display) to count as a *big baby* than an elephant needs to be relative to other elephants to count as a *big elephant*. These results support an account where prototypicality status is taken into account when determining a comparison class in these experiments.

In Experiments 2 and 3, the pattern of prototypicality effects was different from those in Experiments 1a and 1b. When the stimulus images were photographs rather than clipart, prototypically small and neutral objects no longer behaved differently, though these categories were still distinguishable from the category of prototypically big objects, at least some of the time. These differences in prototypicality effects suggest that when the context contains photographic images, and perhaps also in the real world, information about the actual size of the objects depicted is prioritized over prototypicality status information in forming the comparison class. However, when clipart images are judged, prototypicality information is prioritized.

The differences in prototypicality effects emerged when the images in the context of the utterance were changed from clipart images to photographs. Photographs are different from clipart in a number of relevant ways. A photograph depicts a specific individual object. While it may be used to illustrate a category, the depiction is still of a given object, which has particular properties, including dimensions. In contrast,

while clipart images can be used to depict a given individual object, they are also often used to represent a category (Saryazdi et al., 2018; Saryazdi and Chambers, 2018). Since they are not tied to a specific individual object, the depiction often generalizes over the most salient or stereotypical features of a member of the category depicted. The dimensions and proportions do not necessarily correspond to any one individual.

Because of these differences, I propose that when the images in the context are clipart, generalized representations are given more weight in the comparison class, and generic properties of the category are more readily accessed (e.g., that *elephants are big*). The distribution of sizes of objects in the comparison class has a heavier representation of average sized objects, and fewer outliers. On the other hand, when the images viewed are photographs, the comparison class might prioritize representations from the speaker's experience with actual objects of that category. The actual objects experienced have their own dimensions, which are stored in some form in the speaker's object representation. The distribution of sizes in the comparison class might be wider, and less centered around the average. Therefore, while there may be little overlap between the comparison classes of prototypically small and neutral objects when the stimuli are clipart images, there might be more overlap when the stimuli are photographs.

The results of the norming study presented in Chapter 2 are consistent with this hypothesis. I found a significant interaction between image type and the comparison between the size ratings of prototypically small and neutral objects. This interaction is exactly what might be expected if there is a difference in size ratings between prototypically small and neutral objects only when the images are clipart, but not when they are photographs. This result provides further support that prototypicality status becomes a relevant distinguishing factor when there are no actual size differences between the objects of the two prototypicality statuses.

As discussed in Section 3.4, conflicting results are reported in the previous literature investigating differences between clipart and photographs, some finding an effect of image type, others not. For instance, in most tasks in Saryazdi et al. (2018), which norms carefully matched clipart images and photographs, there was no effect of image type on the results. In contrast, in Saryazdi and Chambers (2018), in one

experiment image type had an effect, while in another no such effect was found. In my experiments, as well, there were differences between clipart and photographs. A possible explanation for the conflicting findings in the literature is that the degree of effect image type has on performance may be highly task dependent.

In addition to explaining the confusing picture emerging from the literature, task dependence may also explain why the norming study in Saryazdi et al. (2018) found little difference between clipart and photographs, whereas I found differences between the results of Experiment 1b and Experiment 2. Most of their norming tasks required simply identifying the object depicted. Similarly, in Saryazdi and Chambers (2018), subjects were only required to identify the referent. In my task, subjects were asked about properties of the objects depicted, which may be more dependent on specific features of the image.

Further support for this distinction is provided by looking at the few places Saryazdi et al. (2018) did find differences between clipart and photographs. The two tasks in which significant differences were found were just those tasks that required participants to reason about the objects depicted: familiarity and visual complexity. In the familiarity task participants had to rate how familiar they were with the object in question, and in the visual complexity task participants were asked to rate the image in terms of its intricacy or visual complexity. It is plausible that both of these tasks may depend on specific features of the image. Further work is required to determine the nature of the interaction between task type and image type.

The account presented in this chapter rests on one crucial assumption. The null hypothesis is that *big* and *small* are symmetric in terms of the standard used to determine whether the adjective applies. When provided with a restricted set of items, such as the displays in these experiments, the endpoints of the scale may be restricted to be related to the endpoints in the display. Assuming that the function used to calculate the standard is the same for both adjectives, the zone for each adjective is assumed to be equidistant from the relevant end of the scale. However, there are differences in the structure of the scales associated with the two adjectives, that might offer an alternative explanation. The next chapter investigates this issue further in order to rule out this alternate explanation.

Chapter 4

World Knowledge or Scale Structure Asymmetry?

4.1 Introduction

Chapter 3 presented four experiments examining the nature of the world knowledge integrated with the information from the immediate visual context in determining a comparison class. I argued, based on the reasoning in Tribushinina (2011), that these experiments show that the comparison class used in interpreting dimensional adjectives is influenced by both the information in the immediate visual context and speakers' world knowledge about real-world size information and the prototypicality status of objects.

In all of these experiments, participants are shown displays containing a series of seven images. Each display shows a single object in seven sizes. The size-range of the images in the series is the same across all of the objects. For each series participants were asked "which [objects] are [big/small]?" Across all of the reported experiments, the mean number of images selected as *big* (the *big* zone) is found to be smaller than the mean number of images selected as *small* (the *small* zone). This result is predicted if speakers are integrating information from the immediate visual context with prior world knowledge. If speakers were to only rely on the immediate context, then the expected pattern would be symmetric *big* and *small* zones, whereas if they were to

only rely on world knowledge then all of the images are expected to count as *small*, since all of the images are smaller than the objects they depict.

The observed asymmetry favoring the *small* zone supports an account in which world knowledge is integrated with the information from the immediate visual context in order to determine the comparison class. I will call this the INTEGRATION HYPOTHESIS. However, there is another possible explanation for this asymmetry. It is possible that the asymmetry is a result of scale structure differences between the adjectives *big* and *small*. I will call this theory the SCALE STRUCTURE HYPOTHESIS. Both *big* and *small* are considered to be relative adjectives, and therefore in much of the literature they are considered to lexicalize a scale which is open on both ends (e.g., Lehrer, 1985; Kennedy and McNally, 2005).[1] This claim is based on the behavior of these adjectives with maximizing modifiers, as in (12).

(12) a. * The elephant is totally/completely big.

b. * The mouse is totally/completely small.

Maximizing modifiers, such as *completely* or *totally*, are argued to pick out maximal degrees, and therefore are only able to appear with adjectives with an endpoint standard – those claimed by Kennedy and McNally (2005) to lexicalize a scale with a minimal element (a lower closed scale). The fact that negative polarity dimensional adjectives such as *small* are unacceptable with maximizing modifiers, as seen in (12b), is taken to indicate that the scale lexicalized does not have a minimal element, and is, therefore, open on the lower end.

This is not the intuitive interpretation of the size scale. For dimensional adjectives, intuitively, the scale is bounded at the lower end. There is clearly a lower limit for possible size (namely, 0 cm). The standard argument for the counter-intuitive scale-structure of *small* is that the scale is asymptotically bounded at zero. The intuition is that talking about size presupposes that some degree of size is present, so that zero size does not fulfill this presupposition and is thus not relevant, and not on the scale of size. If the property of size is relevant, then the entity possesses some positive quantity of size (Lehrer, 1985, p. 420). Whether the scale of *small* is asymptotically

[1] For a dissenting view, see McNally (2011).

bounded at zero or contains zero as an endpoint, there is an asymmetry between the positive and negative directions of the scale. On the positive side, theoretically, an entity could keep getting infinitely larger. The positive end is truly unbounded. A ball that has a diameter of 1 meter, could be 1 cm bigger, and could keep growing by a centimeter forever. There is theoretically no limit on this possible growth.

This asymmetry in the ends of the scale associated with *big* and *small* suggests a potential alternate account for the results of Experiments 1–3. It might be that speakers are more reluctant to determine an object to be *big* than they are to determine an object to be *small* because they can imagine many possible bigger ones, whereas there is a clear limit on the small end of the scale.[2] It is possible that the *big* zone is smaller than the *small* zone, not because all of the images are small relative to the objects they depict, but rather because there is no limit to how big the images could potentially be, and therefore people are more reluctant to assert that any of them are *big*. The goal of Experiments 4 and 5 is to distinguish between the Scale Structure Hypothesis and the Integration Hypothesis advocated for in the previous chapter.

4.2 Experiment 4: Tiny objects

In order to distinguish between these two possible interpretations of the asymmetry observed in Chapter 3, I conducted an experiment using images that are bigger than the objects they depict. If the Scale Structure Hypothesis is correct, then the results are not predicted to change from previous experiments, as the scale structure of the adjective remains constant. Even with images bigger than the objects they depict, the *big* zone is expected to remain smaller than the *small* zone. However, if the Integration Hypothesis is correct, then the zone size asymmetry is predicted to flip when the real-world size of the objects changes. If the objects depicted are smaller than the images, then if only world knowledge were used to determine the comparison

[2]Even if an object can keep getting smaller infinitely, e.g., by halves, at some point the object gets so small that the distinctions stop being relevant, since it is too small for any practical purpose, or for the naked eye to see.

class, all of the images would be considered *big*. However, since the information from the immediate visual context is integrated with the real-world size knowledge, the asymmetry is predicted to flip so that the *big* zone is **bigger** than the *small* zone.

4.2.1 Methods

Participants

A total of 49 people (38 male, 11 female) who had not participated in any of my other experiments were recruited using MTurk. Self-selection was open to all and participants who were not native speakers of American English were excluded, as discussed in Section 4.2.2.

Materials and Design

The materials for Experiment 4 are similar in layout to the materials used in Experiments 1a, 1b and 2, but instead of using images which are smaller than the objects depicted in them, the images are photographs of very small objects, as in Figure 4.1. Each trial consists of seven images of an object, ranging in size from 38 to 265 pixels wide, increasing by 38 pixels from image to image. Since the experiment is conducted on MTurk, the actual size of the images that participants see varies according to the size of the screen they use, but the relative size of the images proportional to each other remains fixed. The images in each display are arranged in order of size, either ascending or descending from left to right. The objects in the images are: ant, bean, blueberry, candy,[3] earring, fly, hazelnut, ladybug, pill, pin, seed, and tack.[4] Since all of the objects are very small there are no prototypicality status differences between them. The audio question accompanying each display was created using the same text-to-speech generator as in the previous experiments.

[3]An image of a single candy corn piece was used, but the audio used the label "candy".

[4]The image for earring was a small diamond stud, the pin was a bulletin board push pin, and the tack was a brass thumb tack.

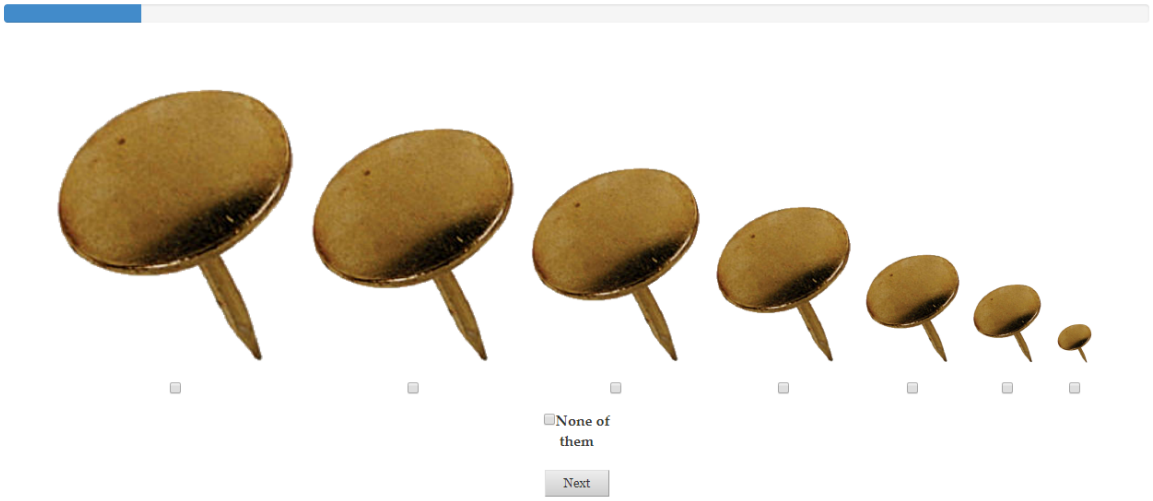


Figure 4.1: Sample trial display for Experiment 4, showing thumb tacks in descending order. The display was accompanied by the audio question "Which tacks are big?"

Procedure

As in Experiments 1–3, two practice trials (as in Experiment 2, using photographs of cars and flowers) are followed by 48 randomized test trials. Each participant sees each object in each order (ascending and descending) with each adjective. The same object never appears twice in a row. As each display appears, the participant hears the prompt, "Which [object]s are [big/small]?" and must select at least one image, or "None of them" before proceeding to the next trial using the "Next" button. The participant cannot advance until the audio prompt finishes. An image can be selected either by clicking on the image itself or by clicking on the checkbox underneath it. Finally, participants were asked to provide demographic information.

4.2.2 Results

Seven subjects were excluded from the analysis based on the exclusion criteria described for Experiment 1a in Section 3.2.1, all for exceeding the threshold for trials with an inappropriate endpoint. Additionally, 29 trials were excluded,[5] leaving a total

[5] 11 trials were excluded for having no endpoint selected, 2 for having the wrong endpoint and 16 for including non-consecutive images.

dataset of 1,987 trials over 42 subjects.

As predicted by the Integration Hypothesis, the zone size pattern flips when the images judged are larger than the actual size of the objects depicted in them. The *big* zone (mean = 3.81, sd = 1.49) was found to be **bigger** than the *small* zone (mean = 1.54, sd = 0.63). A Wilcoxon signed-rank test shows that the difference between the zone sizes is statistically significant ($Z = 21.76, p < 0.001$). An ordinal regression model was constructed with adjective (*big/small*), direction (ascending/descending) and trial number as fixed effects. The maximal random effects structure justified by the design was used. Adjective has a significant negative effect on the number of images selected ($\beta = -8.52, p < 0.001$), indicating that the *big* zone is significantly bigger than the *small* zone. There is no significant effect of direction ($\beta = 0.15, p = 0.138$), or of trial number ($\beta = 0.0007, p = 0.855$). The overall difference between the *big* zone and the *small* zone also holds when looking at individual objects. The *big* zone is clearly larger than the *small* zone for each individual object (Figure 4.2).

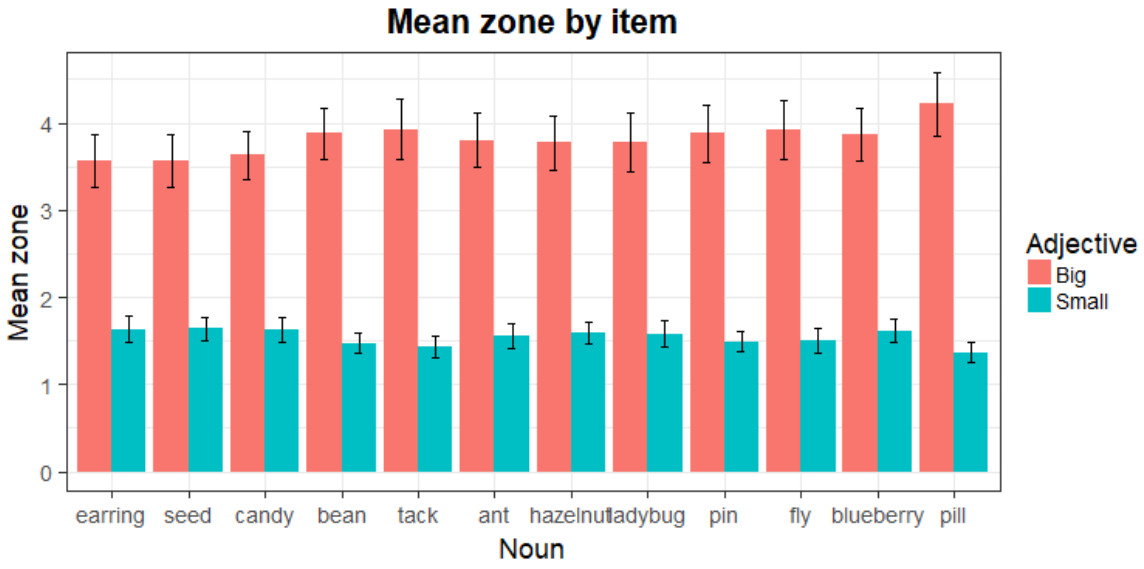


Figure 4.2: The mean *big* (red) and *small* (blue) zones by object in Experiment 4. The error bars represent 95% confidence intervals (CIs).

The stimuli in Experiment 4 were ordered by size, as in Experiments 1–2. As discussed in relation to Experiment 2, when the stimuli are ordered it might be easier

to formulate a response strategy. It would be fairly easy for a participant to simply decide to select some fixed number of images, and answer all of the questions the same without thinking about each individual display. Indeed, as in Experiment 2, a number of participants do not display much variability in their answers, as shown in Figure 4.3. Three participants have completely flat lines, indicating that they selected the same number of images for all of the trials. Another four show very little variability (anon-27, -42, -5 and -9 in Figure 4.3).

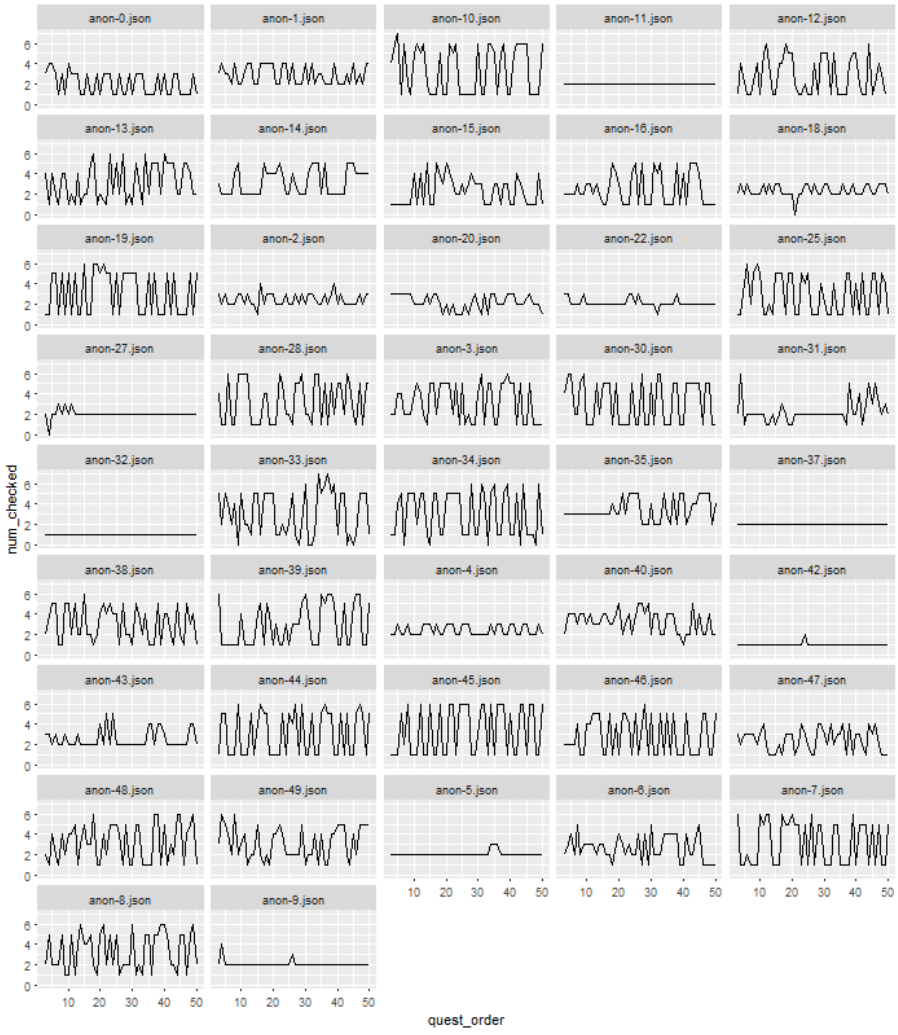


Figure 4.3: Number of images selected by each participant in each trial throughout Experiment 4

4.2.3 Discussion

Experiment 4 demonstrates that when the stimulus images are bigger than the objects they depict, more images are selected as *big* than as *small*, i.e. the *big* zone is bigger than the *small* zone. This is a reversal of the result in Experiments 1–3, in all of which the *big* zone was found to be smaller than the *small* zone. This reversal is predicted by the Integration Hypothesis, and contrary to the prediction of the Scale Structure Hypothesis. Therefore, the Integration Hypothesis provides a better explanation for the zone size effects demonstrated in Tribushinina (2011) and in Experiments 1–3.

If these effects had been caused by the different scale structures of *big* and *small*, then the zone sizes should not have reversed in Experiment 4. The results would have been expected to be the same as in Experiments 1–3, since the scale structure of *big* and *small* does not vary depending on the stimuli viewed. It might have been the case that people were simply more reluctant to call something *big* than *small* since the boundary of how small an object could possibly be is clearer than the boundary of how big it could be. This reluctance would not have changed when the images were exchanged for objects smaller than the images depicting them. The *small* zone would still have been larger than the *big* zone.

However, the Integration Hypothesis predicts that when the images are larger than the objects depicted in them, the *big* zone is bigger than the *small* zone. The knowledge that all of these images are bigger than the objects they depict and, therefore, might count as *big* is integrated with the information from the immediate visual context, providing the boundaries of possible sizes. This would account for the observed asymmetry whereby the *big* zone is bigger than the *small* zone when the images are bigger than the objects they depict.

As discussed in Section 3.4.3, when the stimuli are ordered it might be easier to formulate a response strategy. A participant might decide to select some fixed number of images, and answer all of the questions the same without considering each individual display. Indeed, as in Experiment 2, a number of participants did not display much variability in their answers (Figure 4.3). I did not exclude these participants from the analysis, since it is possible that they did consider each display and arrive at this pattern. However, to address the concern that the results are simply

attributable to response strategies, Experiment 5 randomizes the order of the images in the display so as to make it harder to form a strategy for answering, and thus ensure that the results observed in Experiment 4 are reliable.

4.3 Experiment 5: Randomized tiny objects

Experiment 5 is identical to Experiment 4, except that the order of the images is randomized in each display, as it was in Experiment 3 (see Section 3.5). The goal of the randomization is to make it harder for participants to form a response strategy. Additionally, the randomized order is more naturalistic, as objects discussed and experienced in the world are rarely organized by size.

4.3.1 Methods

Participants

A total of 111 people (62 male, 49 female) who had not participated in any of the previous experiments were recruited using MTurk.[6] Self-selection was open to all and participants who were not native speakers of American English were excluded, as discussed in Section 4.3.2.

Materials and Design

The materials for Experiment 5 are similar in layout to Experiment 3, but similar in stimuli to Experiment 4. The images are photographs of very small objects displayed in randomized order, as seen in Figure 4.4. Each trial consists of seven images of an object, ranging in size from 38 to 265 pixels wide, increasing by 38 pixels from image to image. Since the experiment was conducted on MTurk, the actual size of the images that participants see varies according to the size of the screen they use, but the size of the images relative to each other remains fixed. The images appear in

[6]Since the stimuli in this experiment appeared in a randomized order, there was no factor of direction of display and therefore, half the number of trials as in Experiment 4. Double the number of people as in Experiment 4 were recruited in order to achieve the same power.

a row in randomized order. As in Experiment 4, the objects depicted in the images are: ant, bean, blueberry, candy, earring, fly, hazelnut, ladybug, pill, pin, seed, and tack. Since all of the objects are very small, there are again no prototypicality status differences between them. The audio question accompanying each display was created using the same text-to-speech generator as in the previous experiments.



Figure 4.4: Sample trial display for Experiment 5, showing thumb tacks in randomized order. The display was accompanied by the audio question "Which tacks are big?"

Procedure

As in Experiments 1–4, there are two introductory trials (identical to those in Experiments 2–4, using photographs of cars and flowers), followed by 24 test trials. Each participant sees each object with each adjective. The trials appear in a randomized order for each participant, with the constraint that the same object not appear twice in a row. As each display appears, the participant hears the audio prompt, and then is required to select at least one image, the "None of them" box or the "All of them" box, before proceeding to the next trial using the "Next" button. The participant cannot advance until the audio prompt finishes. An image can be selected either by clicking on the image itself or by clicking on the checkbox underneath it. Finally, as in the previous experiments, participants were asked to provide demographic information.

4.3.2 Results

Fourteen participants were excluded from the analysis based on the exclusion criteria described for Experiment 1a (Section 3.2.1); two were excluded for using a screen smaller than 12 inches, four for reporting a native language other than English, five for exceeding the threshold of marking non-consecutive images, and three for exceeding the threshold of trials with no endpoint. Additionally, 39 trials were excluded,[7] leaving a total dataset of 2,289 trials over 97 subjects.

As predicted by the Integration Hypothesis, and as observed in Experiment 4, the sizes of the zones flip when the images in the context are larger than the real-world size of the objects depicted. The *big* zone (mean = 3.49, sd = 1.16) was found to be **bigger** than the *small* zone (mean = 1.96, sd = 0.73). A Wilcoxon signed-rank test shows that the difference between the zones is statistically significant ($Z = 21.94, p < 0.001$). The difference in size between the zones is slightly smaller than in Experiment 4, perhaps as a result of the extra noise introduced by the more complex task, where participants may also have had to mentally reorder the stimuli in order to perform the task. The overall difference between the *big* zone and the *small* zone also holds when looking at individual objects, as seen in Figure 4.5.

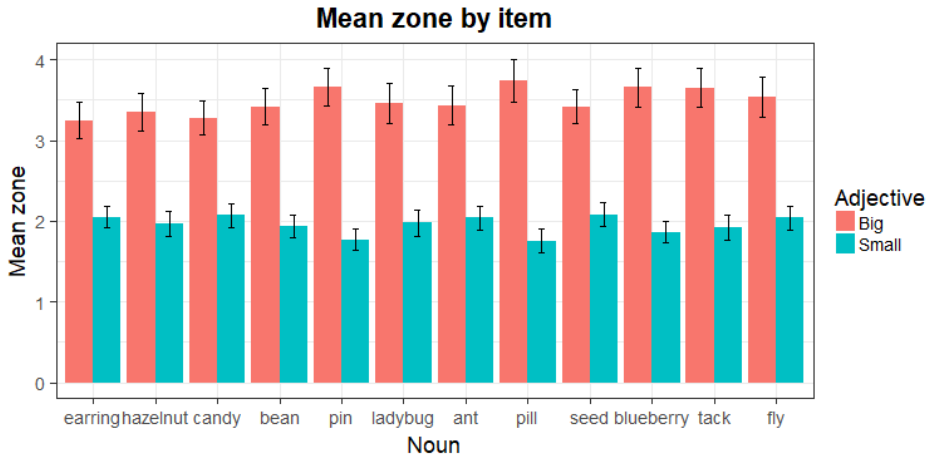


Figure 4.5: The mean *big* (red) and *small* (blue) zones by object in Experiment 5. Error bars represent 95% CIs.

[7] 17 trials were excluded for not providing a good endpoint and 22 trials were excluded for not marking consecutive images.

An ordinal regression model was constructed with adjective (*big/small*) and trial number as fixed effects. The maximal random effects structure justified by the design was used. Adjective has a significant negative effect on the number of images selected ($\beta = -6.44, p < 0.001$), indicating that the *big* zone is significantly bigger than the *small* zone. There is no significant effect of trial number ($\beta = -0.002, p = 0.689$).

Unlike in Experiment 3, there does seem to be a reduction in the number of participants answering the experiment strategically (Figure 4.6).

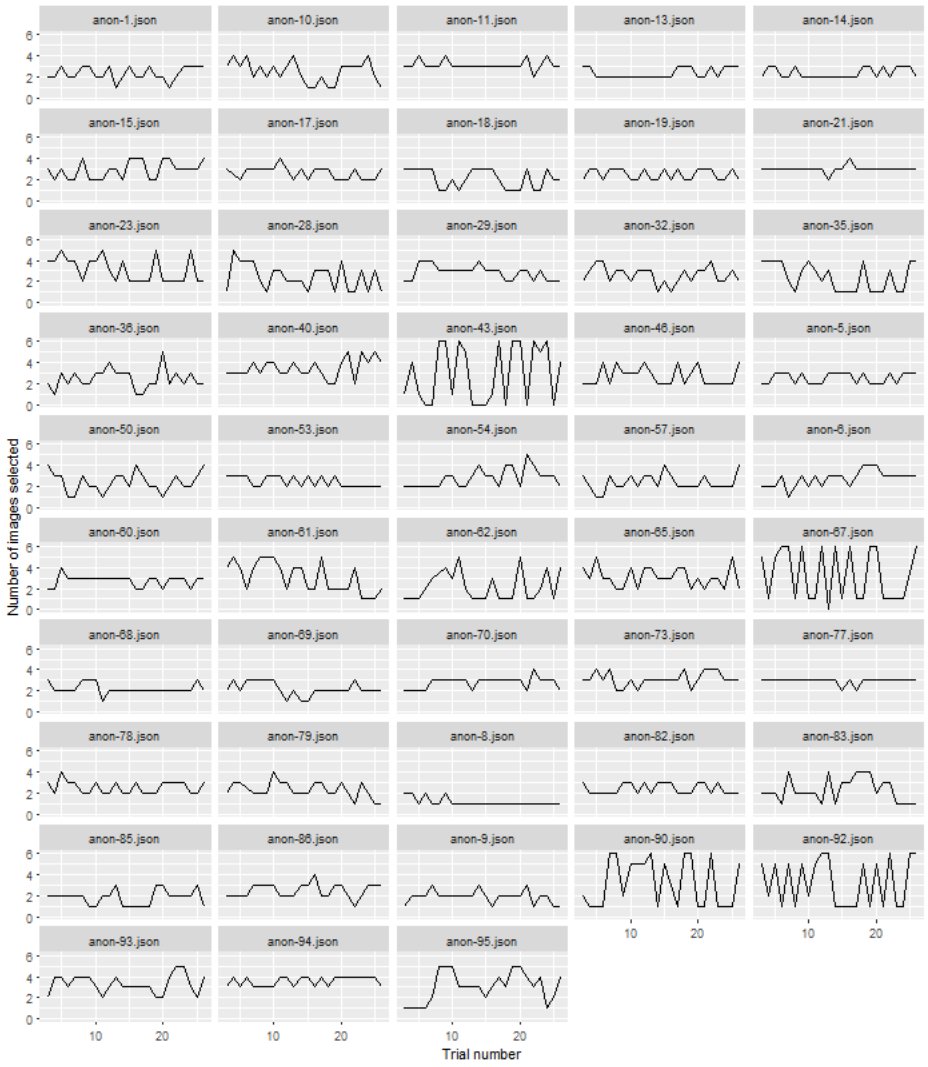


Figure 4.6a: Number of images selected by each participant in each trial throughout Experiment 5 for the first half of the participants

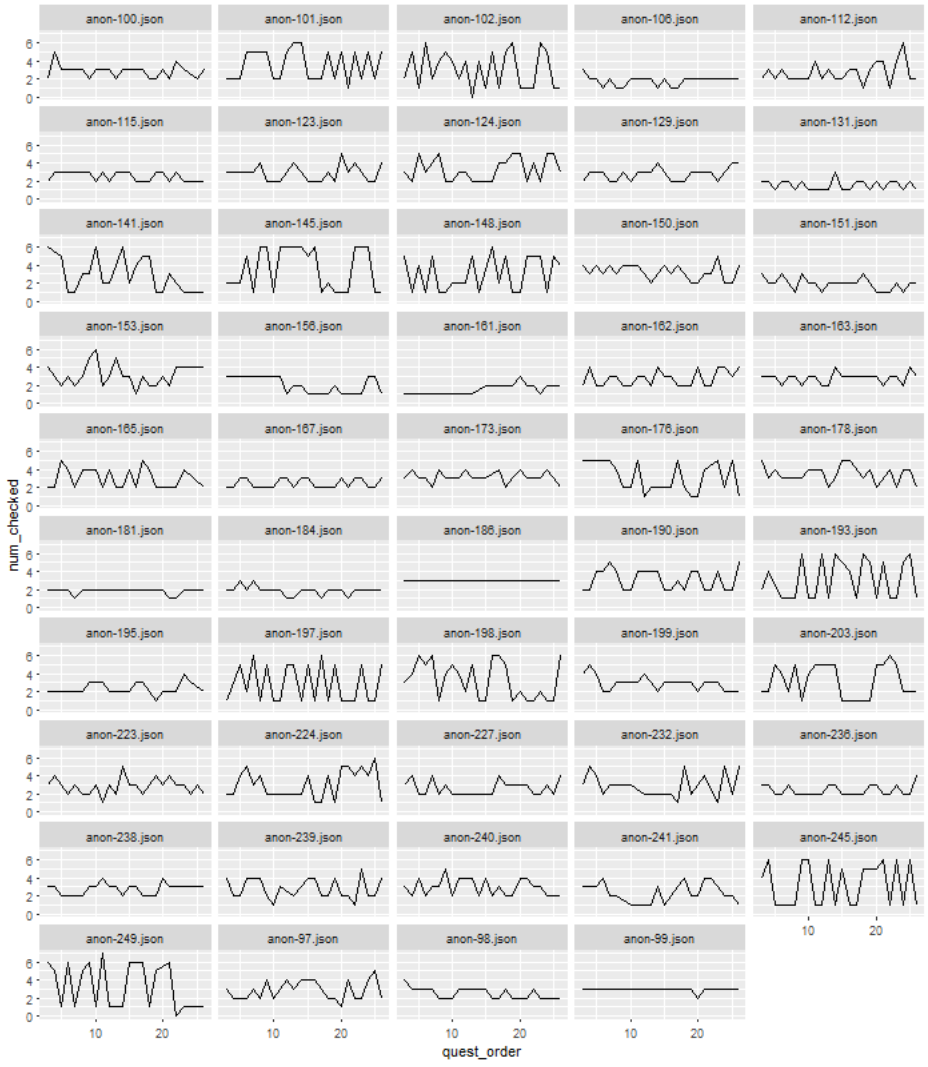


Figure 4.6b: Number of images selected by each participant in each trial throughout Experiment 5 for the other half of the participants

There is only one participant who does not vary at all in the number of images they selected over the course of the experiment (anon-185, Figure 4.6b), and fewer who settle in to a fixed response pattern over the course of the experiment.

4.3.3 Discussion

In Experiment 5, the *big* zone was found to be bigger than the *small* zone, as in Experiment 4. The difference between the zone sizes is highly significant in Experiment 5, though the effect is weaker than in Experiment 4. The randomization of the order of the stimuli adds noise to the process, but not enough to change the effect. The bigger *big* zone found in Experiment 5 provides further support for the Integration Hypothesis, and argues that the results from Experiment 4 are not merely attributable to a particular strategy selected by a participant. When the order of the images is randomized on the screen, an additional step is required in order to apply a response strategy, since first the stimuli must be ordered, therefore encouraging participants to think about each display as it is presented and answer the accompanying question.

4.4 General Discussion

The *big* zone was found to be smaller than the *small* zone in Experiments 1–3. This result was predicted under the Integration Hypothesis, the hypothesis that speakers integrate world knowledge about the size of objects with information from the immediate visual context when determining the comparison class used in interpreting relative adjectives. However, this result would also be predicted by the Scale Structure Hypothesis. The different scale structures of *big* and *small*, whereby *big* is unbounded on the upper end of the scale, but *small* is bounded on the lower end may also cause such an asymmetry in zone sizes. Since it is possible to know how far the object being judged is from the endpoint of the scale for *small*, but not for *big*, speakers might be more comfortable judging an object as the former than the latter.

Experiments 4 and 5 argue against this interpretation. The Scale Structure Hypothesis depends on the adjective being judged rather than the objects, and therefore does not predict that there should be a difference in the size of the zones based on the size of the objects in the context. The Integration Hypothesis, on the other hand, predicts that if the images in the context are bigger than the objects they depict, then the size of the zones relative to each other should be reversed, so that the *big*

zone is **bigger** than the *small* zone. This reversed effect is exactly what was observed in Experiments 4 and 5.

The persistence of this result in Experiment 5 argues that the reversed effect observed in Experiment 4 is not caused by undue use of response strategies. The effect persists when the order of the images in each display is randomized, making it harder for participants to answer by rote, based on a predetermined strategy.

Not only did the overall size of the *big* and *small* zones flip in Experiments 4–5 in comparison to the results from Experiments 1–3, the difference in size between the *big* zone and the *small* zone is also larger. Speculatively, the zone size might also influenced by the images being fairly close to the real-world size of the objects depicted in them. The smallest image participants saw is only marginally bigger than the real-world size of the object depicted in it, if at all, and the biggest is seven times larger than that. In comparison, even the biggest image of the prototypically small objects in Experiments 1–3 (baby, duckling, bunny, gnome) is significantly smaller than the real-world size of the object depicted in it. This difference is exaggerated even more for the prototypically big objects. A real elephant is orders of magnitude bigger than even the biggest photograph of an elephant seen by participants.

The size of the images of tiny objects is closer to the real-world size of the objects depicted than is the case for any of the other images. This resemblance to reality may have encouraged participants to place prioritize their real-world knowledge over the information from the comparison class when determining which images counted as *big* or *small*. Since all of the images are bigger than the objects depicted in them, if world knowledge were the only information being used to determine the comparison class, then all of the images would have counted as *big*. If they were compared to a class made up of experiences with real objects, such as real beans, the bean images are all in the top $n\%$, or above average, and thus would count as *big*. Counting more images as *big* than were counted as *small* in Experiments 1–3 may indicate an emphasis on knowledge of real-world size. From Experiments 4 and 5, along with the difference in results between Experiments 1 and 2, it seems plausible to posit that the more the stimuli resemble the real-world objects they are meant to represent, the more prior world knowledge regarding the size of these objects in the real world comes into play.

Chapter 5

The Structured Acquisition of Dimensional Adjective Antonyms

5.1 Introduction

In the first part of this dissertation, I zoomed in on a specific facet of adjective meaning and interpretation, namely the comparison class. I focused on one pair of dimensional adjectives, *big* and *small*. Experiments probing the adult interpretation of this pair of adjectives provided insight into the composition of the comparison class, the nature of world knowledge and the role of context in the interpretation of these adjectives. In this chapter, I zoom out to look at the broader semantic field of dimensional adjectives, and take a different perspective on questions of context in relative adjective interpretation. In this chapter, I investigate the acquisition of the semantic field, demonstrating the importance of contrast in anchoring children's production of dimensional adjectives. I also investigate the gradual process of acquiring this semantic field, and how the structure of the acquisition process reflects the structure of the semantic field.

Dimensional adjectives such as *big* and *long* are among the first adjectives acquired by children. These adjectives form a highly structured semantic domain, or semantic field, consisting of related antonym pairs, such as *wide* and *narrow*. Each word in this semantic field must be acquired, as must the relations between the words. Previous

work on the acquisition of this semantic field has focused on English speaking children in a fairly small age range, who are no longer at the beginning of their use of dimensional adjectives. Therefore, this work has not been able to provide a detailed enough view of the course of acquisition to be able to fully capture the process of acquiring this semantic field. In an experiment eliciting dimensional adjective antonyms from Hebrew speaking children covering a broader range of ages than previous work, I show that the structured nature of the semantic field is reflected in the acquisition process.

Section 5.2 presents theoretical background on the structure of the semantic field of dimensional adjectives, antonymy, Hebrew dimensional adjectives and the acquisition of dimensional adjectives. This section lays the theoretical foundation for the experiment described in Section 5.3. Section 5.4 concludes the chapter.

5.2 Background

Before describing the experiment, the next sections introduce the necessary background. Section 5.2.1 introduces the lexical semantics of dimensional adjectives. Core to the semantics of dimensional adjectives, and providing structure to the semantic field is antonymy, described in Section 5.2.2. As the experiment was conducted in Hebrew, Section 5.2.3 introduces Hebrew dimensional adjectives and compares them to English ones. Finally, Section 5.2.4 discusses previous work on the acquisition of the semantic field and provides the framework within which my results are discussed and interpreted.

5.2.1 Dimensional Adjective Meaning

Dimensional adjectives have a number of components which make up their meaning. They are associated with a scale of degrees of a property relating to extension in space, as opposed to color (like *red*) or appearance (like *ugly*). Most dimensional adjectives also specify in which direction this extension occurs (e.g., vertically for *tall* or horizontally for *long*), although *big* and *small*, for example, relate to size generally without specifying a particular dimension. Each adjective denotes extension towards one end,

or pole, of the scale, and thus has either positive or negative polarity ([±pol]). Each adjective is also related by antonymy to an adjective denoting extension towards the other end of the scale (e.g., Lyons, 1977; Cruse, 1986). Antonymy will be explored in more detail in Section 5.2.2. These meaning components of dimensional adjectives are summarized in Table 5.1 with examples of the components for a number of dimensional adjectives.

Adjective	Dimensionality	Direction	polarity	antonym
tall	✓	+vertical	+pol	*short*
narrow	✓	−vertical, +secondary	−pol	*wide*

Table 5.1: Examples of the meaning compononents of dimensional adjectives

Bierwisch (1967) provides a framework for understanding how dimensional adjectives specify the direction they describe. He analyzes a dimensional adjective as a cluster of features which encode the dimensions that the adjective picks out. For example, *big/small* apply to an n-space, where *n* can refer to one, two or three dimensions depending on what object the adjective is being applied to. Terms like *tall/short* are more specific, and refer to one particular dimension, [1-space]; more specifically they can be thought of as encoding the property [+vertical], as can *high/low*. The other dimensional adjective antonyms refer to non-vertical dimensions. *Long/short* can be characterized as [1-space, −vertical], while *wide/narrow, thick/thin* and *deep/shallow* are [1-space, −vertical, +secondary], where secondary refers to a dimension that is not the most salient dimension of the object to which the adjective is applied. For example, the most salient dimension of a door might be its height, then its width and finally its thickness. The more features an adjective specifies, the more cognitively complex it is (Clark, 1972). Thus, *big* and *small* [n-space] are the least complex of the dimensional adjectives since they do not specify any direction. Given the features just described, the dimensional adjectives form the complexity hierarchy in (13) where *big/small* is less complex than *tall/short, high/low* and *long/short*. Among the non-vertical adjectives, *long/short* is less complex than *wide/narrow, thick/thin* and *deep/shallow*, which all specify an additional feature [+secondary]. The theory does not distinguish among these latter three pairs in terms of complexity.

(13) a. big/small < {tall/short, high/low, long/short}

b. long/short < {wide/narrow, thick/thin, deep/shallow}

As discussed above, the members of antonymous pairs are differentiated by a polarity feature, [±pol]. Since antonymy is one of the major organizing features of dimensional adjectives, it is discussed in more depth in Section 5.2.2.

5.2.2 Antonymy

Both members of an antonym pair share most meaning components, differing only with regard to polarity. Thus, in a sense, antonymy relies on the other meaning components of an adjective, since the other components must be shared. However, sharing a scale and differing in polarity is not enough to make two adjectives good antonyms. For example, *dry* is a good antonym of *wet*, while near synonyms of *dry*, such as *arid* or *parched*, are not. Two adjectives which are good antonyms must also have significant overlap in the semantic categories of nouns they are used to modify. For example, while *dry* and *sweet* might be considered antonyms when discussing wine, there are many uses of *sweet* which cannot be contrasted with *dry* (for an extensive discussion on the properties of antonyms, see Muehleisen, 1997). Additionally, antonyms often appear in the same syntactic frames, and contrast with each other in the same contexts, as in the sentence in (14) (Justeson and Katz, 1991; Jones, 2007; Muehleisen, 1997).

(14) That was one more reason she didn't look forward to Cathy's visit, short or long. (Justeson and Katz, 1991, p.10)

Thus, an adjective's antonym is a lexical property that must be learned from observing how that adjective is used in relation to other lexical items. For example, one cannot learn that *large* is the antonym of *small* but not of *little* (Muehleisen, 1997) simply from observing uses of *large*. It is also necessary to keep track of the nouns each adjective is used to modify, as well as the other adjectives appearing in the same contexts.

Within a pair of antonyms, the member with positive polarity ([+pol]) is typically considered unmarked, and the member with negative polarity ([−pol]) marked (e.g.,

Greenberg, 1966; Lehrer, 1985; Lyons, 1977; Zwicky, 1978). The most general criterion for determining markedness is that the unmarked adjective is neutralized in questions and nominalizations. In (15a) *tall* is not biased about which part of the scale is involved, whereas, in (15b), *short* is biased, implying that Jane is short.

(15) a. How tall is Jane?

b. How short is Jane?

The unmarked member of an antonym pair can also appear in more contexts than the marked one. For example, *5 feet tall* is normal, while *5 feet short* is odd. Perhaps as a result of the less restricted contexts, the unmarked member is also more frequent (Greenberg, 1966; Lehrer, 1985; Zwicky, 1978).[1] These difference may have consequences for the acquisition of these adjectives, discussed in Section 5.2.4.

5.2.3 Hebrew Dimensional Adjectives

Like English, Hebrew has an open adjective class with a mostly similar inventory of dimensional adjectives (Tribushinina et al., 2014). The biggest difference in the adjective classes in the two languages is that Hebrew has a fairly diverse class of modal adjectives, where other languages use modal verbs or inflections (Ravid and Nir, 2000). Turning to dimensional adjectives, there are two main differences between the Hebrew inventory of these adjectives and the English one. Whereas English uses *short* as the antonym of both *long* and *tall*, Hebrew has two separate words, *kacar* for the former and *namux* for the latter (henceforth translated as *short-l* and *short-t*). Conversely, where English distinguishes between *high/low* and *tall/short*, Hebrew (similar to other languages, such as German), does not, using *gavo'a/namux* for both meanings. There may be subtle lexical-semantic distinctions, but in general the meanings of the other Hebrew dimensional adjectives map straightforwardly to those found in English.

While the semantics of dimensional adjectives is similar in English and in Hebrew, all other aspects differ. Phonologically the words are obviously different. They are of different lengths, and made up of different phonemes. Syntactically, in Hebrew, adjectives follow the noun they are modifying as opposed to in English where

[1] For more complete list of markedness properties of antonym pairs, see Lehrer (1985, p.400).

they precede the noun. Morphologically, adjectives in Hebrew agree in gender (masculine/feminine) and number (singular/plural[2]) with the noun they are modifying, indicated by suffixes on the adjective, and also agree in definiteness by means of the definite article clitic. The addition of this morphological material sometimes leads to phonological changes in the adjective stem as well, as seen in the example of the adjective paradigm in Table 5.2.

The citation form of the adjective is the masculine singular form. Number agreement starts to appear in children's speech before gender agreement, masculine agreement before feminine. Both of these appear in subject-verb agreement before they are found within the noun phrase, perhaps because children only start using a wide range of adjectives around age 3;0, after they consistently produce sentences with subjects and verbs (Berman, 1985; Levy, 1983a). Since all nouns participate in the binary gender system, and gender is pervasive across the system, children acquire the system fairly early on, and make relatively few errors with gender (Berman, 1985; Levy, 1983b). Children generally can assign gender agreement from nouns to adjectives by around ages 3;0–3;6, and have more trouble with the feminine forms than with the (unmarked) masculine, in both singular and plural, initially adding a masculine plural marker to a feminine noun, or using a masculine verb form despite its subject being feminine. (Berman, 1985).

Masc. sg.	Masc. pl.	Fem. sg.	Fem. pl.
pil gadol 'big elephant'	pilim gdolim 'big elephants'	kubiya gdola 'big block'	kubiyot gdolot 'big blocks'

Table 5.2: The Hebrew adjective paradigm for *gadol* ('big')

5.2.4 The Acquisition of Dimensional Adjectives

Moving away from Hebrew specifically, children generally start to produce adjectives somewhat later than nouns and verbs and do not use them very frequently

[2]There is also a dual form, but it is marginally productive, and is not relevant to the current study.

until the age of 1;8 (Stolt et al., 2008). By age 3;0 adjective frequency reaches a plateau (Tribushinina and Gillis, 2012). Among the first adjectives to be acquired cross-linguistically are color, size, and physical properties (a rather diverse category including adjectives denoting properties relating to surface, smell, taste, sound, shape, temperature, light, consistency and speed) (Tribushinina et al., 2014). Looking across many semantic categories of adjectives in five languages (Turkish, Hebrew, German, French and Dutch), Tribushinina et al. (2014) determine that during the initial period of adjective use, dimensional adjectives, and particularly *big* and *small* (or their counterparts in other languages), are prominent in the speech of both parents and their children.

One of the things children must learn when beginning to use adjectives is which ones form antonym pairs with each other. As discussed in Section 5.2.2, one of the properties of antonymous adjectives is that they appear in contexts and syntactic frames emphasizing the contrast between the two adjectives. Corpus work has shown that children are exposed to many such contrastive uses of dimensional adjectives and, starting at least as early as age 2;0, they use them in these contexts themselves (Jones, 2002; Jones and Murphy, 2005; Murphy and Jones, 2008). The contrast may in fact ease the acquisition of the meaning of these terms (Tribushinina et al., 2013). Tribushinina et al. (2013) show that across 16 children acquiring eight different languages there is a correlation between the amount of contrast heard and increase in adjective production. Hebrew is not one of the eight languages they look at in this study,[3] but there is no reason to think the discourse function of contrast and the benefits it provides for acquisition are significantly different in Hebrew. As mentioned in Section 5.2.2, in order for two adjectives to be good antonyms it is not sufficient that they relate to opposite ends of a shared scale. Additionally, the two adjectives must have significant overlap in the semantic categories of nouns they are used to describe. In order to acquire antonyms, a child must, therefore, track not only the contexts in which an adjective is used with regard to the situation in the world that it is describing, but also which other adjectives have been used in such situations.

[3]The languages they examined are Croatian, Dutch, French, German, Italian, Lithuanian, Russian and Turkish.

One question that has been much discussed in the literature on the acquisition of antonyms is whether the two members of an antonym pair are acquired together or whether there are significant differences in the age of acquisition between them. The negative, marked member of the pair, which is also less frequent, is generally posited to be more complex and therefore harder to acquire than the positive member (Barner and Snedeker, 2008; Clark, 1971, 1973; Donaldson and Wales, 1970; Ehri, 1976; Smith et al., 1986; Townsend, 1976). Other studies have found similar acquisition patterns for both members of the antonym pair (e.g., Clark, 1972).

The acquisition of the semantic field of dimensional adjectives is a structured process. Clark (1972) argues that the order of acquisition of the dimensional adjectives is driven by their semantic complexity, as defined in Bierwisch (1967). Clark also uses the domain of dimensional adjectives to investigate the process of acquiring a semantic field. She conducted an elicitation experiment examining eight pairs of dimensional adjectives: *big/small, tall/short, high/low, long/short, old/young, wide/narrow, thick/thin* and *deep/shallow*. Based on the responses children provided in an antonym elicitation task, and on the answers they provided that were not adult-like, she argues that the acquisition of these adjectives is gradual, and that children associate certain words with others in the same semantic field even before they have acquired the full meaning of these words.

Clark formulates order of acquisition predictions based on Bierwisch (1967). As stated in Section 5.2.1, Bierwisch characterizes *big/small* as [n-space] terms and therefore the least complex. *Tall/short* and *high/low* are both [1-space, +vertical]. *Long/short* is [1-space, −vertical]. These three pairs are indistinguishable in complexity in terms of the theory. Among the non-vertical terms, *long/short* is less complex than *wide/narrow, thick/thin* and *deep/shallow*, which are all [1-space, −vertical, +secondary]. The final pair, *old/young* is less clearly placed in this field, but can be thought of at least as [1-space] because it places things on the temporal axis. From this hierarchy, Clark generates the order of acquisition predictions spelled out in (16), arguing that complexity plays a major role in determining the order of acquisition.

(16) a. big/small < {tall/short, high/low, long/short}

b. long/short < {wide/narrow, thick/thin, deep/shallow}

Clark (1972) uses a word game to investigate these order of acquisition predictions. She asks children between the ages of 4;0 and 5;5 to provide the antonyms of a list of words. They are given an example, "if *I* say *happy*, *you* say ... sad", followed by dimensional adjectives in the same sentence template, where they need to provide the antonyms ("If *I* say *tall*, *you* say ..."). In order to discourage a default strategy of simply applying syntactic negation (e.g., giving the antonym of *tall* as *not tall*), in a second experiment, children hear the list of adjectives embedded within three-word full noun phrases (e.g., *a tall man*). Clark (1972) coded children's responses as being Adult Correct (AC) if they provided the expected adult antonym, and as Semantically Appropriate (SA) as long as the response had different polarity than the term in the prompt and shared a semantic feature (all AC responses were also SA). Thus, *little*[4] in response to *high* is coded as SA. The task at hand was an abstract word game, inaccessible to children under the age of 4. The abstract nature of the task also made it more difficult when the word was not well known to the child as there was very little contextual support. Clark found that both SA and AC responses increased significantly with age.

Clark posited that the percent of correct responses provided to the adjectives reflects the order of acquisition. Adjectives with fewer correct responses are acquired later. When the adjectives are ordered by the percent of correct responses, the order, Clark found, reflects the order of acquisition predicted based on the complexity hierarchy (17).

(17) a. AC: big, tall, long, old, high, thick, deep, wide

b. SA: big, tall, long, high, old, thick, wide, deep

Under this theory the more specific the adjective in terms of which dimensions it encodes, the more semantically complex, and hence more cognitively complex it is. More complex adjectives should be acquired later than less complex ones. In addition to generating order of acquisition predictions, the complexity hierarchy also allows Clark to make predictions about the substitutions children make when they do

[4]Clark (1972) treats *little* and *small* as equally appropriate antonyms for *big* and combines both of these responses in her results. She does not discuss potential differences between these adjectives in her paper.

not provide what is considered the AC antonym. Namely, since children acquire the less complex terms first, they are predicted to only substitute less complex adjectives for more complex ones, and not the reverse. She found that *big/small*, the least complex adjectives, accounted for 80% of the substitutions for dimensional terms. Other, more complex, terms are never substituted for *big* and *small*. She argues that this substitution pattern demonstrates that children start to construct semantic fields before they have completely acquired the meaning of the words making up the field, associating the adjective heard with other dimensional adjectives of the appropriate polarity, and therefore making an appropriate substitution, despite not being able to provide the AC antonym.

Clark (1972) studied children well beyond the initial stages of adjective acquisition and looked at English. Follow-up work looking at younger children focused on comprehension rather than production, so that the tasks would be suitable for the age of the participants (Bartlett, 1976; Brewer and Stone, 1975; Eilers et al., 1974; Klatzky et al., 1973). However, while comprehension tasks are informative about certain aspects of lexical development, they do not provide information about the acquisition of the association between the two members of an antonym pair. They can, therefore, be performed correctly without complete acquisition of the lexical entries in question.

Thus far, elicitation of antonyms with children under the age of four has not been succesful, and it remains to be shown that the early stages of acquisition are consistent with the data in Clark (1972). Since corpus work (e.g., Murphy and Jones, 2008[5]) has shown that children as young as two already use members of antonym pairs in contrast with each other, they should be able to produce these antonyms in an experimental setting as well at this age. In my study, I show that when adjectives are presented with the proper context, accompanied by images of objects contrasting in the dimension mentioned, rather than in an abstract word game, children are able to produce antonyms earlier than previously demonstrated. Data from younger participants than those previously examined will help provide a more complete picture

[5]This study looked at a wide range of antonym pairs which among others, included dimensional adjectives.

of the acquisition process and shed light onto the process of constructing lexical entries and building up the semantic field of dimensional adjectives. Not only is it helpful to have a window into what these younger children already know, but even their incorrect responses can be illuminating. When their lexical knowledge is less comprehensive, children develop strategies to cope with not yet having acquired the complete adult meaning of the adjectives, further demonstrating the gradual nature of the acquisition process.

The same order of acquisition predictions should hold for younger children, so that *big/little* is expected to precede the other dimensional adjective pairs (see (17) for the detailed order predictions), and may be substituted for them before the more complex adjectives have been acquired. Substituting an incorrect adjective that has the right polarity and comes from the right semantic field suggests that the children associate various dimensional adjectives with each other as belonging to the same semantic field. Such substitutions also suggest that even before children have a complete representation of the meaning of the adjective and its antonym, they know the polarity of the word.

The complexity hypothesis presented in Clark (1972) should not be language specific, and should apply to any language with similar lexical semantics of dimensional adjectives. The order of acquisition of dimensional adjectives, as well as the substitution patterns found, should therefore be the same in all such languages. If other factors, such as syntactic, phonological, or morphological complexity played a more significant role in driving the order of acquisition, then one might expect the order to differ in significant ways in languages which differ from English in those respects. Therefore, conducting similar studies in other languages provides a good test of the validity of the complexity hypothesis.

Therefore, I ran an experiment, eliciting dimensional adjective antonyms from Hebrew speaking children starting at age three. This experiment is described in Section 5.3.

5.3 Antonym Acquisition Experiment

In order to gain insight into the process of dimensional adjective acquisition, it is important to study children as close as possible to the beginning of their adjective use. I predict that if the elicitation task is more concrete than previous tasks used, anchored to pictures of objects rather than being an abstract word game, then younger children would also be able to successfully provide antonyms. Having more context provided in the task may also make it easier for young children, as it has been shown that contrast helps children acquire and understand language.[6] Therefore, I devised a task eliciting dimensional antonyms within a context which provides visual, prosodic and syntactic contrast.

5.3.1 Methods

This study uses a more concrete antonym elicitation task than the one used in Clark (1972) to enable children under the age of four to participate. Children were shown pictures exemplifying the dimensional contrast between two otherwise identical objects, and provided with the description of one of the pictures. They had to provide the description of the other. The study was conducted in Hebrew rather than English in order to provide further, cross-linguistic support for semantic complexity being a driving force in the order of acquisition of dimensional adjectives.

Participants

The participants in the experiment were 28 children between the ages 3;2 and 8;10. One additional child, age 7;8, was excluded since she got bored halfway through the

[6]In addition to the work on adjective acquisition previously discussed, contrast has been shown to be helpful in the acquisition of nouns (e.g., Au and Markman, 1987). Another linguistic domain in which providing contrast has helped children perform better on a task is scalar implicature. Barner et al. (2011) show that making children aware of the relevant scalar alternatives improves their performance on a scalar implicature task, rejecting sentences such as *some of the horses jumped over the fence*, when in fact all of the horses did. They argue that this demonstrates that children's difficulty with scalar implicature is due to difficulty in deriving the correct alternative sets, not to lack of pragmatic sophistication. Providing the appropriate contrasts for children helps them demonstrate more adult-like competence on the implicature task.

task and started providing both members of the antonym pairs, but did not always provide the intended adjectives. The children were all native speakers of Hebrew born and living in Israel (although 9 were bilingual, speaking both Hebrew and English), and the testing was conducted in Hebrew. They were recruited through friends and friends of friends. The children were all asked if they would like to read a story with the experimenter and the parents all provided informed consent.

A pilot study with three two-year-olds showed that they were too young for the task. The three children produced a total of 29 responses (instead of the 42 expected), due to a short attention span. They did, however, provide evidence of being able to produce the contrast between *gadol* ('big') and *katan* ('small'). Additionally, before they lost patience, they provided some interesting errors instead of the adult-correct, more complex dimensional adjectives. One type of error made was to produce a color or shape adjective in response to the dimensional stimuli. Such an answer provides suggestive evidence that while they realize that the word heard and the answer required are adjectives, they have not yet situated the prompt word within the semantic field of dimensional adjectives.

Materials

Seven pairs of dimensional adjective antonyms were tested, the Hebrew translations of the adjectives used by Clark (1972).[7] These adjectives appear in the second and third columns of Table 5.3. Four different stimulus sets were constructed for the seven pairs of antonyms. Each set started with the same three introductory images using non-dimensional adjectives. These were intended to familiarize the children with the task. Following the introduction, the set of seven target adjectives was presented, followed by another repetition of the same adjectives in a different order and with a different picture, so that each child heard each adjective twice. Two of the stimulus sets contained one member of each pair of antonymous adjectives, and the other two stimulus sets had the complementary set of adjectives. Each set was made up of 17

[7] *Old* and *young* are not strictly dimensional adjectives. She characterizes them as 1-space by virtue of their being situated on the temporal axis, and thus includes them in the study (Clark, 1972). I left them in for comparability to the original study.

scenes (one per page), each of which contained a picture of Elmo and a picture of Cookie Monster, each in possession of an object, identical but for the dimension being tested. For example, for the antonym pair *big/small*, Elmo might have a big block and Cookie Monster a small block, as in the sample scene in Figure 5.1. The objects depicted for each adjective are listed in the two rightmost columns of Table 5.3.

Status	Positive	Negative	Picture 1	Picture 2
Intro	naki ('clean')	meluxlax ('dirty')	shirt	N/A
Intro	male ('full')	rek ('empty')	cup	N/A
Intro	patuax ('open')	sagur ('closed')	door	N/A
Target	gadol('big')	katan ('small')	block	elephant
Target	arox ('long')	kacar ('short')	pencil	belt
Target	gavoa ('tall/high')	namux ('short/low')	building	orange
Target	aveh ('thick')	dak ('thin')	marker	notebook
Target	zaken ('old')	cair ('young')	man	woman
Target	raxav ('wide')	car ('narrow')	window	door
Target	amok ('deep')	radud ('shallow')	sandbox	pool

Table 5.3: The adjectives tested and the nouns depicted for each item. Picture 1 appeared in the first half of the stimulus set and Picture 2 in the second half.

The stimulus sets were constructed by randomizing the set of adjectives twice, making up the two halves of the set. Polarities were assigned in a random order to the first half, and then preserved in the second half, so that each child saw only one member of an antonym pair. The two halves were then flipped for the second stimulus set using those polarities, so that the first order became the second and vice versa. The same was done for the complementary members of the antonym sets. Each child was randomly assigned to one of the four stimulus sets.

Procedure

The children were tested in private homes, in a quiet room. In most cases, only I was present during the testing, although with a few of the younger children, the mother was also present, but was asked not to interfere. The stimuli were presented on an

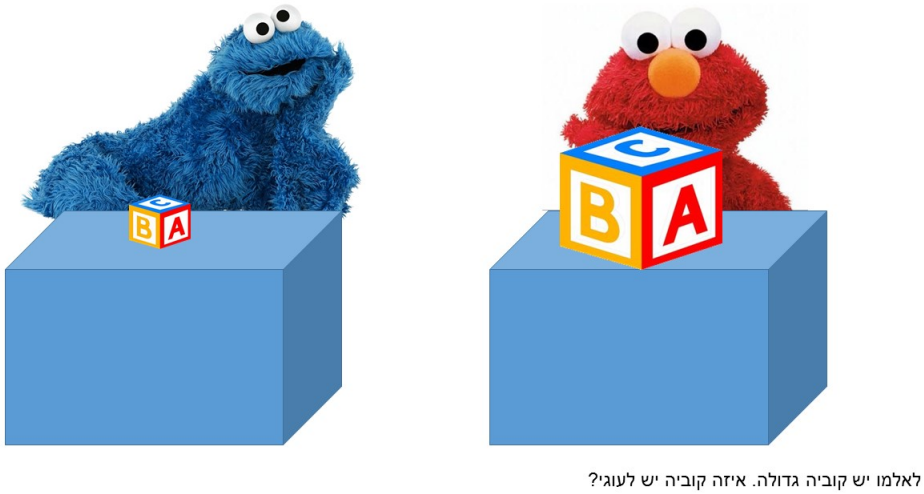


Figure 5.1: A sample stimulus scene for the adjective pair *big/small*, where the target for elicitation was *small*.

iPad, and the whole session was recorded using an audio recorder. The children were introduced to the characters, Elmo and Cookie Monster, and walked through the first three pictures. If they needed help on these, it was provided. Then the target images were presented, one at a time. For each image, the child was read a prompt such as the example in (18), which was intended to elicit the antonym (in this case *small*) from the child.

(18) li'elmo yeš kubiya gdola. Eyze kubiya yeš li'ugiflecet?
DAT.Elmo exists block big. which block exists DAT.Cookie-Monster?
'Elmo has a big block. Which block does Cookie Monster have?'

Positive feedback (e.g., 'right' or 'great') was provided after each response, regardless of what it was. If no response was provided the child was prompted once more. Responses of 'don't know' were not given a second chance and no corrections were made during the task. The answers provided were transcribed from the audio recording. In cases where the child gave more than one answer, only the first was counted for the purposes of analysis.

5.3.2 Results

The children's responses to the prompts such as the one in (18) fell into five categories, as set out in (19), based on the categorization used in Clark (1972).

(19) a. AC responses – the antonyms provided in Table 5.3

b. SA responses – responses with correct polarity which share semantic features with the AC response. In addition to including all the AC responses, these also include three other categories of response:

i. Negation (Neg) – the negation of the stimulus, e.g., *lo zaken* ('not old') in response to *old*

ii. Substitution of *small/big* (SB) – *katan* ('small') or *gadol* ('big') substituted for the AC response (with the correct polarity)

iii. Other – a near synonym of the AC response, for example *anak* ('giant') in response to *katan* ('small')

c. Other Dimensions (OthDim) – a response with the correct polarity but the wrong dimension, for example, *arox* ('long') in response to *dak* ('thin')

d. Error (Err) – a response with inappropriate polarity or not a dimensional adjective (e.g., color or shape)

e. Don't know (DK) – 'I don't know' or no response at all.

Response Strategies

The distribution of responses in each category is provided in Table 5.4. Overall, across all adjectives and across all age groups, 77.4% of the responses were semantically appropriate, showing that children are fairly good at providing an appropriate response even when they do not yet know the standard antonymous adjective.

Response Type		Percent
SA (Total: 77.4%)	AC	50.4
	Neg	9.4
	SB	16.9
	Other	3.5
OthDim		9.1
Error		5.6
DK		5.1

Table 5.4: Distribution of response types

Age effects

When the responses are examined by age of the child, as shown in Figure 5.2, there is a Pearson's correlation of .79 between age and percent of AC responses ($t = 6.47$, $df = 25$, $p < .05$). However, the Pearson's correlation of .19 between age and percent of SA responses is not significant ($t = .96$, $df = 25$, $p = .3475$). By age three, age is not a significant predictor of the percentage of SA responses a child produces. It seems that by this age children minimally have some familiarity with most of the terms and have strategies to cope with the lack of access to the AC form, and thus provide an SA response most of the time. However, while they succeed in the communicative task of conveying the appropriate meaning, they do not yet have a good grasp of the correct antonyms. The only adjective pairs that the three-year-old participants provided any AC responses to were *gadol/katan* ('big/small') and *arox/kacar* ('long/short').

Differences in response patterns by age group

As children get older, not only do they respond correctly to more adjectives, but their response strategy when they do not know the correct answer changes. For this analysis, I divided the children into four age groups, three-year-olds ($n = 6$), four-year-olds ($n = 6$), five- and six-year-olds ($n = 8$) and seven- and eight-year-olds ($n = 8$). Figure 5.3 shows the percent of responses of each type provided by each age group. Errors, which were observed frequently among the two-year-olds responses in

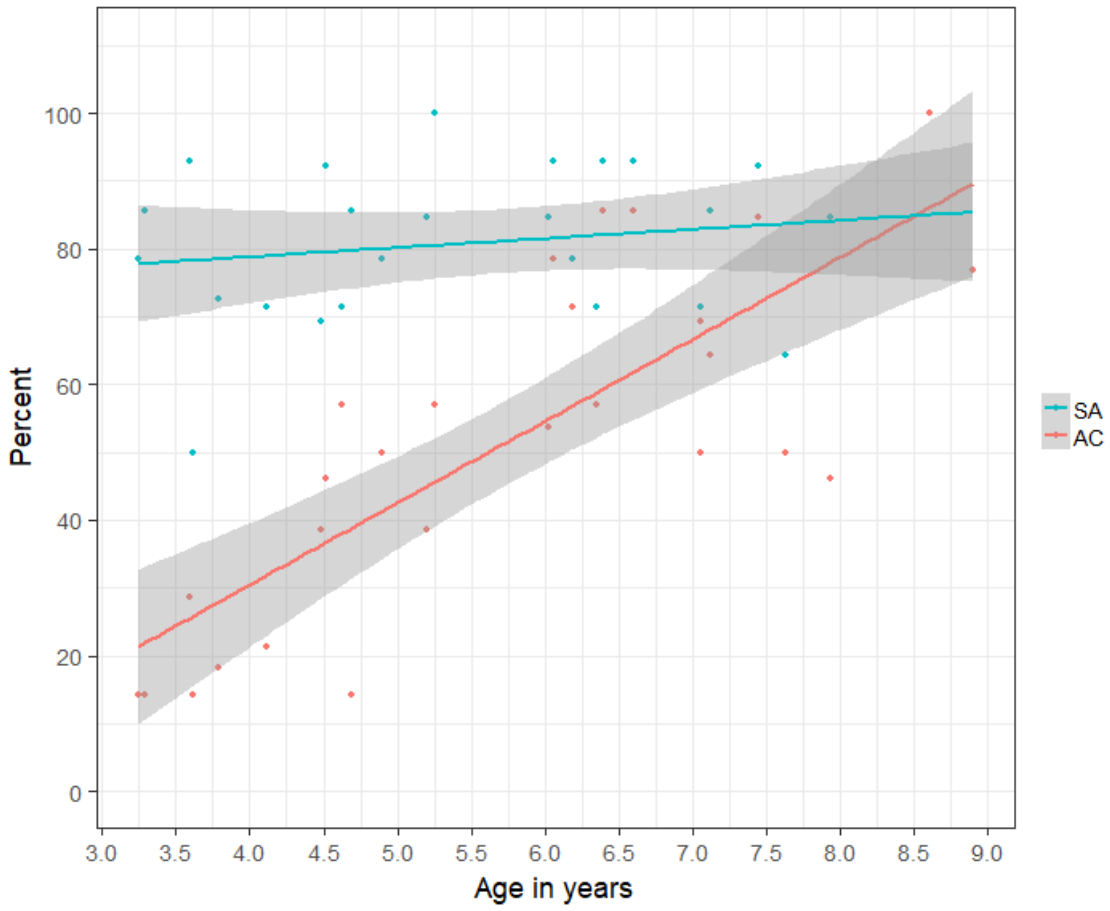


Figure 5.2: The percent of SA (blue) or AC (red) responses for each participant by their age in years

the pilot study, are still present but less prevalent for the three- and four-year-olds. Errors are rare for the older children, suggesting that, by age five, children solidify their knowledge of the semantic field and familiarity with the adjectives involved. Use of *katan/gadol* ('small/big') rather than the AC response is found in all of the age groups, but is most popular as a strategy for the three-year-olds; after that age it makes up an increasingly small percent of the responses. Negation is also used by all age groups, but makes up the largest percent of the responses of the four-year-olds, and this strategy becomes less frequent after age five. It is important to note that each age group contains a small number of children, and therefore these patterns are

only suggestive. A larger study, with more children in each age group, is needed in order to reliably draw conclusions as to the chronological development of the different response strategies. The results of my study provide a foundation which can be used to inform such a study.

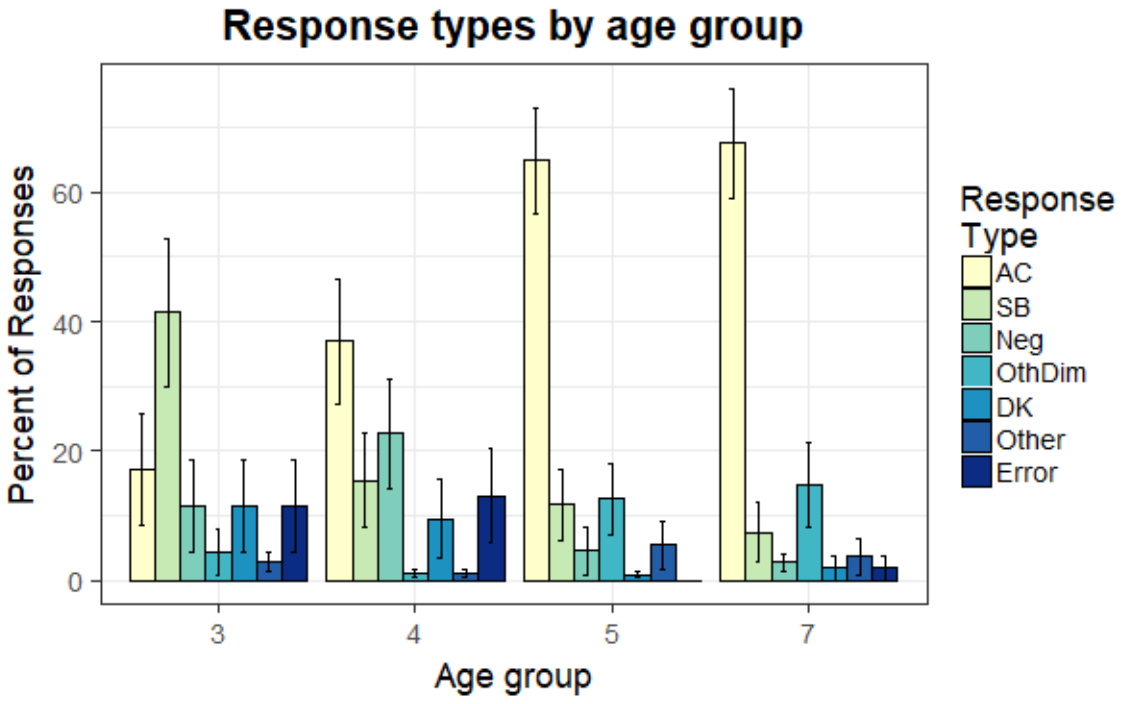


Figure 5.3: Distribution of response types for each age group. The error bars represent 95% confidence intervals (CIs) calculated by bootstrap sampling from the distribution of response types for each age group 10,000 times.

Order of Acquisition

Since each child saw each adjective twice, a score was assigned to each adjective for each subject: 1 if they provided two AC responses, .5 if they provided one AC response, and 0 for no AC responses. Similarly, an SA score was also calculated based on the number of SA responses each child provided. I found no differences between the scores of positive and negative members of each of the antonym pairs, similar to the result reported in Clark (1972). Therefore, for the analysis of the order of

acquisition of the adjectives, I will combine the data for the two members of each pair, and use the positive adjective to refer to this combined data.[8] The AC and SA scores were independently averaged for each adjective pair across the subjects (Table 5.5, Figure 5.4).

	Adult Correct					Semantically Appropriate				
Adjective	3	4	5	7	Mean	3	4	5	7	Mean
big/small	1	1	0.94	0.94	0.96	1	1	0.94	1	0.98
high-tall/low-short	0	0.42	1	0.81	0.63	0.7	1	1	1	0.94
long/short	0.2	0.67	0.63	0.75	0.59	1	1	1	1	1
thick/thin	0	0.33	0.81	0.94	0.59	0.5	0.75	0.88	0.94	0.80
old/young	0	0.08	0.56	0.81	0.43	0.7	1	1	1	0.94
deep/shallow	0	0.08	0.56	0.25	0.26	0.5	0.33	0.88	0.44	0.56
wide/narrow	0	0	0.06	0.25	0.09	0.7	0.25	0.38	0.31	0.39

Table 5.5: Mean scores by age group for each adjective pair

Tukey HSD tests (which correct for the problems arising from running multiple t-tests) were conducted to determine significant differences between the pairs. The AC scores present a slightly complex picture. *Gadol* ('big') is significantly different from all of the other pairs. *Gavoa* ('high/tall'), *aveh* ('thick'), and *arox* ('long') are not different from each other, but are different from *amok* ('deep') and *raxav* ('wide'). *Zaken* ('old') is only different from *gadol* on the one side and *raxav* on the other. In other words, the average score for *zaken* is significantly smaller than *gadol*, and significantly higher than *raxav*, but not significantly different from any of the other adjectives tested. These results are summarized in (20), where those terms underlined together are not significantly different from each other at a $p < .05$ level. The SA results in (21) are simpler, showing two groups of adjectives: *raxav* ('wide') and *amok* ('deep') in one group, and all of the other adjectives in another. The mean scores of *raxav* ('wide') and *amok* ('deep') are significantly lower than the scores of the rest of the adjective pairs, but not significantly different from each other. The scores of

[8]There were, nevertheless, some indications of differences in the pattern of responses between the positive and negative adjectives, which will be discussed below in the section discussing positive vs. negative adjectives.

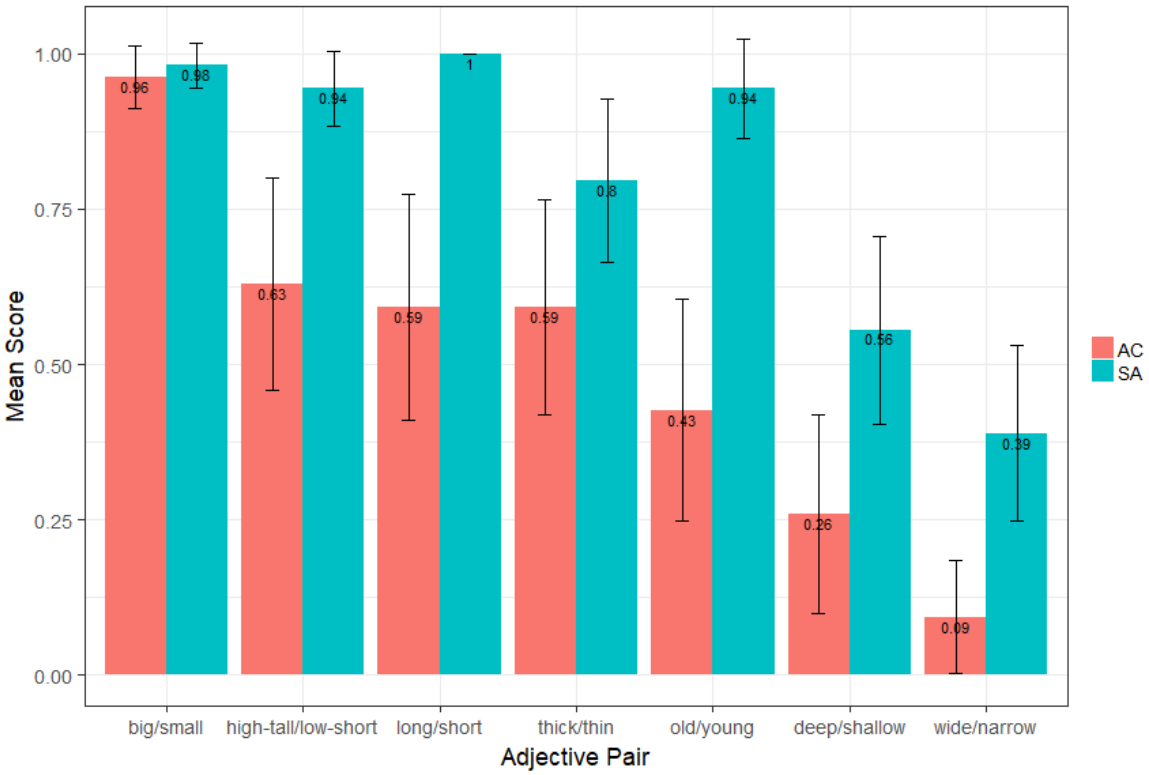


Figure 5.4: Mean AC (red) and SA (blue) scores for each adjective pair collapsed across age groups. Error bars represent 95% CIs.

the rest of the adjectives also did not differ significantly from each other. Again, the average scores of adjectives underlined together did not differ significantly from each other.

(20) AC: big, high/tall, long, thick, old, deep, wide

(21) SA: long, big, high/tall, old, thick, deep, wide

The results in (20) and (21) bear out the predictions made in Clark (1972) regarding the order of acquisition ((22)–(23), equivalent to (17), with the notation changed so that it is consistent with the reporting of the results), with one exception regarding the placement of *aveh* ('thick'). The performance on *aveh* was not different from the performance on *arox* ('long'), although the latter was predicted to be less complex

and acquired first. According to the complexity hierarchy based on Bierwisch (1967) (see Section 5.2.4), *long* is [1-space, −vertical], while *thick* is [1-space, −vertical, +secondary]. Since *thick* further specifies that it refers to a secondary dimension, it is more complex. However, while no difference between *arox* and *aveh* was found, crucially, the directionality of the inconsistency with the predictions was not reversed. In no case was an adjective pair that was predicted to be harder than another one found to be significantly easier than it.

(22) AC: big, tall, long, old, high, thick, deep, wide

(23) SA: big, tall, long, high, old, thick, wide, deep

Setting aside this one exception of *aveh*, the rest of the predictions were borne out. Finding a similar order of acquisition in Hebrew as was found for English supports the theory that the semantic complexity of the dimensional adjectives is driving the order of acquisition. There are no obvious differences in the semantics of these dimensional adjectives between Hebrew and English. However, if the driving factor in the order of acquisition had been some properties of the phonology or morphology of the adjectives, one might have expected to find much more significant differences between languages in the order of acquisition.

Substitutions

Two types of substitutions are common among the children's responses to the stimuli. The first type is substitution of *gadol/katan* ('big'/'small') for other, more specific dimensional terms, while preserving the correct polarity of the answer (SB, see (19b-ii)). This type made up 16.9% of the total responses. The second common type of substitution was substitution of a different dimension of the correct polarity, which made up 9.1% of the total responses (OthDim, (19c)). When the polarity was incorrect the response was scored as an error. SB substitutions were scored as semantically appropriate, while OthDim ones were not. All substitutions which occurred more than once appear in Tables 5.6 and 5.7.

Both types of substitutions provide insight into how the children build up their lexical network of dimensional terms. Substitution of *katan* and *gadol* for the other

expected → substitution	*gadol* 'big'	*arox* 'long'	*gavoa* 'high/tall'	*zaken* 'old'	*aveh* 'thick'	*raxav* 'wide'	*amok* 'deep'
gadol 'big'		9	8	5	3	7	
katan 'small'							2
arox 'long'					2	5	
aveh 'thick'						7	
ribua 'square'						1	
xalak 'smooth'							2
mevugar 'grown up'				5			
total	0	9	8	10	5	20	4

Table 5.6: Substitutions occuring more than once for positive polarity adjectives. The horizontal axis is the expected AC response, and the vertical axis is the word substituted. The table contains the number of times that each substitution occurred.

expected → substitution	*katan* 'small'	*kacar* 'short'	*namux* 'low/short'	*cair* 'young'	*dak* 'thin'	*car* 'narrow'	*radud* 'shallow'
gadol 'big'			2			1	2
katan 'small'		13	6		2	4	6
kacar 'short'			1		3	4	2
arox 'long'						1	
namux 'low/short'	1						5
aveh 'thick'					1		
dak 'thin'						4	
ribua 'square'						1	
malben 'rectangle'						2	
yeled 'boy'				3			
total	1	13	9	3	6	17	15

Table 5.7: Substitutions occuring more than once for negative polarity adjectives. The horizontal axis is the expected AC response, and the vertical axis is the word substituted. The table contains the number of times that each substitution occurred.

dimensional adjectives suggests that the child knows that the term they heard was also a dimensional adjective, as is the response required, but has not yet learned what the appropriate antonym is.[9] The pattern of substitutions conforms to the predictions made about complexity and order of acquisition. While *katan* and *gadol* were often

[9]Since the children are responding correctly to some of the adjectives while providing SB responses to others, it is not likely that they do not understand the task.

substituted for more specific dimensions, it was much less common to find the reverse. The hardest adjectives, *raxav/car* ('wide/narrow') and *amok/radud* ('deep/shallow'), were never substituted for other terms, although other adjectives were frequently substituted for them. The younger children occasionally provided an adjective that was not even dimensional. These errors were very rare for the older children, which indicates that, while they may not yet know the appropriate antonym, they do know that they are supposed to provide a dimensional term. It is unlikely that the younger children's non-dimensional errors indicate complete lack of understanding of the task, since those same children could provide an SA response for other adjectives. However, it is possible that they indicate distraction or other performance difficulties. Another possible cause for such answers is not associating the adjective heard with dimensional terms, and therefore not providing even an SA response.

The response pattern suggests that the children start constructing the semantic field of dimensional adjectives in their mental lexicon before they have fully fleshed out lexical entries for these adjectives. One of the first features they have in place seems to be the polarity of the adjective. It was exceedingly rare for a child to provide a term of the wrong polarity. The pictures may have also helped with this task, as the object they were asked about clearly differed in size from the object they heard described.

Positive vs. Negative Adjectives

I found no significant differences in performance between positive and negative members of any of the antonym pairs, as determined using the same scores as were used for calculating the order of acquisition, described in Section 5.3.2. In other words, when the average score for the positive member of the antonym pair was compared to the average score of the negative member of the antonym pair, the difference between the two scores was not found to be significant for any of the pairs. The lack of difference is consistent with the finding in Clark (1972), where no difference was found between the positive and negative members of each antonym pair.

However, there do seem to be some systematic differences between the positive and negative members of a pair when looking at other measures. The data suggest

that there might be a difference in age of acquisition between positive and negative adjectives. One way to crudely capture age of acquisition is to look at how old a child is when they first produce a word. My data is not particularly well suited for looking at this measure, since I only have a single sample from each child at a single point in time, and no other information about each child's language use history. Additionally, each child only saw either the positive member of an antonym pair or the negative member. I do not have data regarding how they would respond to the other member of the pair. Finally, the ages are not equally balanced between the two groups. However, despite these fundamental issues, it is interesting to note that there is an almost entirely consistent pattern regarding age of use in my sample. The youngest child in my sample to provide an AC response to any given negative adjective was always younger than the youngest child to provide an AC response to the corresponding positive member of the pair, with one exception. The youngest child to provide an AC response to *ave* ('thick') was younger than the youngest child to provide an AC response to *dak* ('thin'), as illustrated in Figure 5.5.[10] I speculate that the differences in youngest age of first response may correspond to actual differences in age of acquisition between the positive and negative members of the antonym pairs.

In light of the issues discussed in the previous paragraph, the age of first mention in the experiment is not a perfect measure of age of acquisition. However, the youngest child to provide an AC response did vary among the adjectives, showing that the pattern is not the result of a single precocious child answering everything correctly long before their peers. Additionally, as demonstrated by the order of the pairs as presented in Figure 5.5, the age of the youngest child to provide an AC response increases in roughly the same order as the order of acquisition shown in (20), providing some evidence that this measure may be informative.

A paired t-test shows that the mean age of first mention is significantly different for positive adjectives than for negative adjectives ($t = 2.397, df = 7, p = 0.048$). In order to provide a correct response to a prompt containing the negative adjective, the child must associate the negative adjective that they heard with the positive one that

[10]For this analysis I am including the two-year-olds who successfully answered the prompt since I am interested here in the youngest age of correct response. Since I am looking at each adjective pair individually, the partial data from the two-year-olds poses less of an issue.

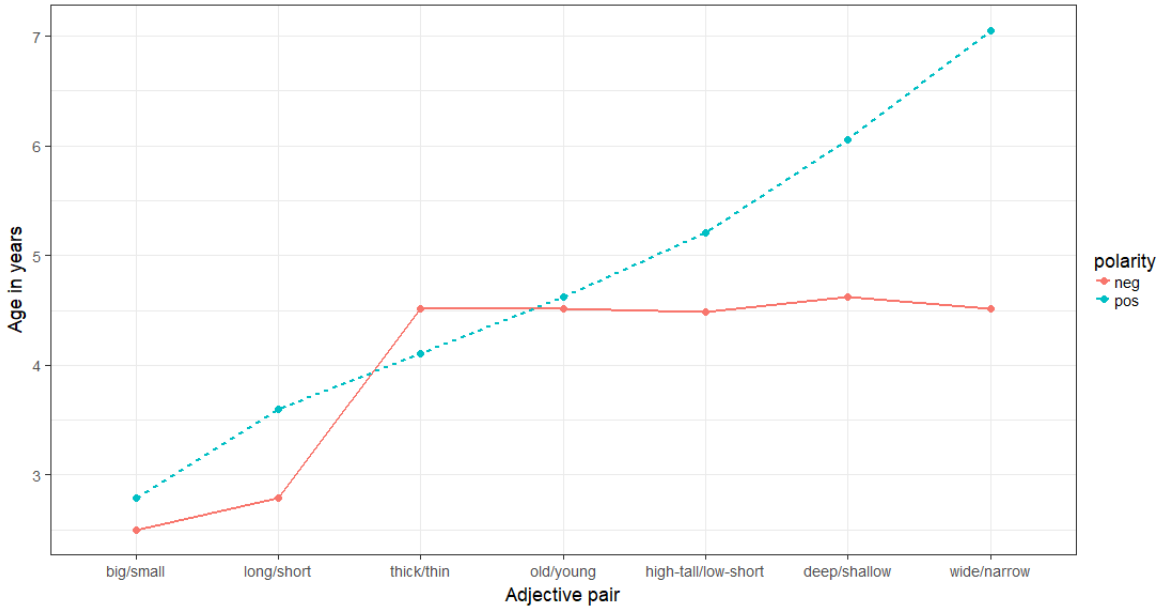


Figure 5.5: Age in years of the youngest child to provide an AC response to each negative (solid red line) and positive (dashed blue line) adjective. Note that the line for the negative adjectives remains almost exclusively below the line for the positive adjectives.

they must produce and then produce the positive adjective. The positive adjective is the more frequent (Clark, 1972) of each pair, as well as being unmarked. The consistently younger age at which the AC response is first provided for the negative adjectives may indicate that it is easier to produce the positive form than the negative form. This result is consistent with other work finding asymmetries between the positive and negative members of antonym pairs, in which children had more trouble with the negative member of the pairs (Barner and Snedeker, 2008; Clark, 1971, 1973; Donaldson and Wales, 1970; Ehri, 1976; Smith et al., 1986; Townsend, 1976).

Another difference between the positive and negative members of an antonym pair is the number of unique responses provided. While each prompt had one AC response, participants could, in principle, provide any word they wanted in response. Thus, an adjective whose antonym was known by all the participants would receive only one unique response – the AC antonym. On the other hand, an adjective whose antonym was less familiar might receive a number of different responses. In this experiment,

in addition to some adjectives having more different responses than others, there are also differences in the number of different responses positive and negative members of any given antonym pair received.

This measure also gives some indication of the difficulty of an adjective. The adjectives that children had more trouble with and seem to be acquired later received more unique responses than the ones acquired earlier. Thus, *big* and *small* had very little variability in responses, while many different answers were provided for *wide* and *shallow*. This is also true within pairs. When there is a substantial difference in number of unique responses it is the positive, unmarked member which has more unique responses than the negative, marked member, as shown in Figure 5.6.

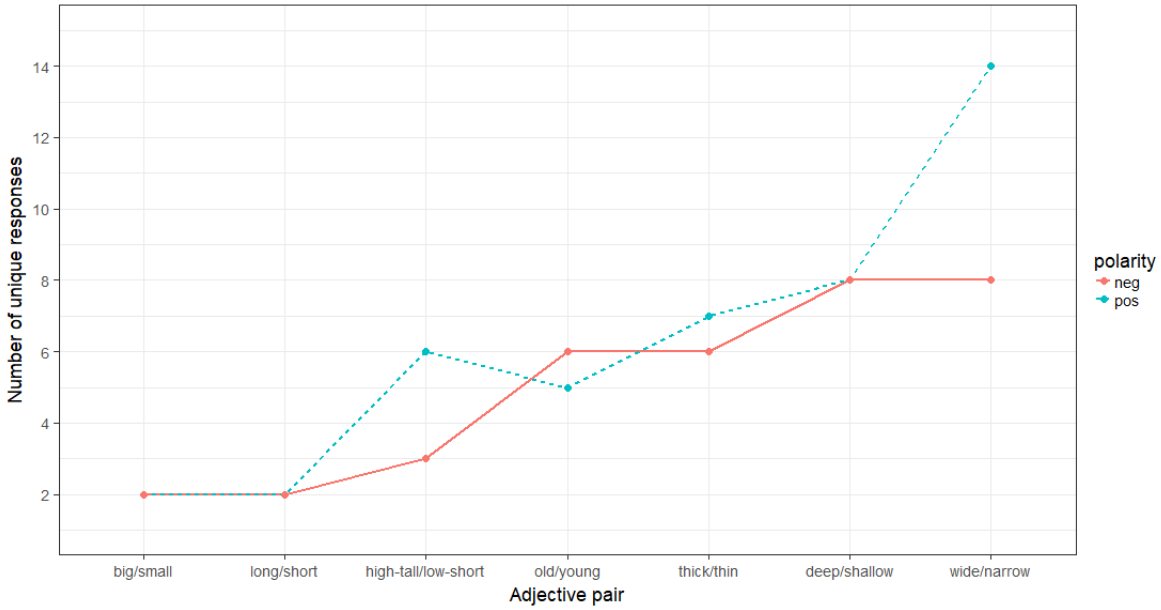


Figure 5.6: Number of unique responses to each negative (solid red line) and positive (dashed blue line) adjective. Note that the line for the negative adjectives for the most part remains equal to or below the line for the positive adjectives, indicating fewer unique responses.

A lower number of unique responses may indicate a stronger association with the correct response. This measure serves as an additional indication that producing the negative member to a positive prompt is harder than producing the positive one and, therefore, when required to produce the negative member of a pair, across participants

there was more variability in the responses. On the other hand, when the negative member of an antonym pair appears in the prompt, the participant needed to produce the positive member of the pair. The positive member of the pair is unmarked and more frequent. It appears more often without the negative member than the negative member appears in a context without the positive member (Muehleisen, 1997). There were also, for the most part, fewer unique responses to the negative member than to the positive member. This may indicate that the positive member is easier to produce, or more strongly associated with the negative member of the antonym pair than vice versa. This point is further strengthened if the 29 responses of the two-year-olds are included, and then for every pair except *zaken* ('old') and *gadol* ('big') the positive member has more unique responses than the negative (Figure 5.7).

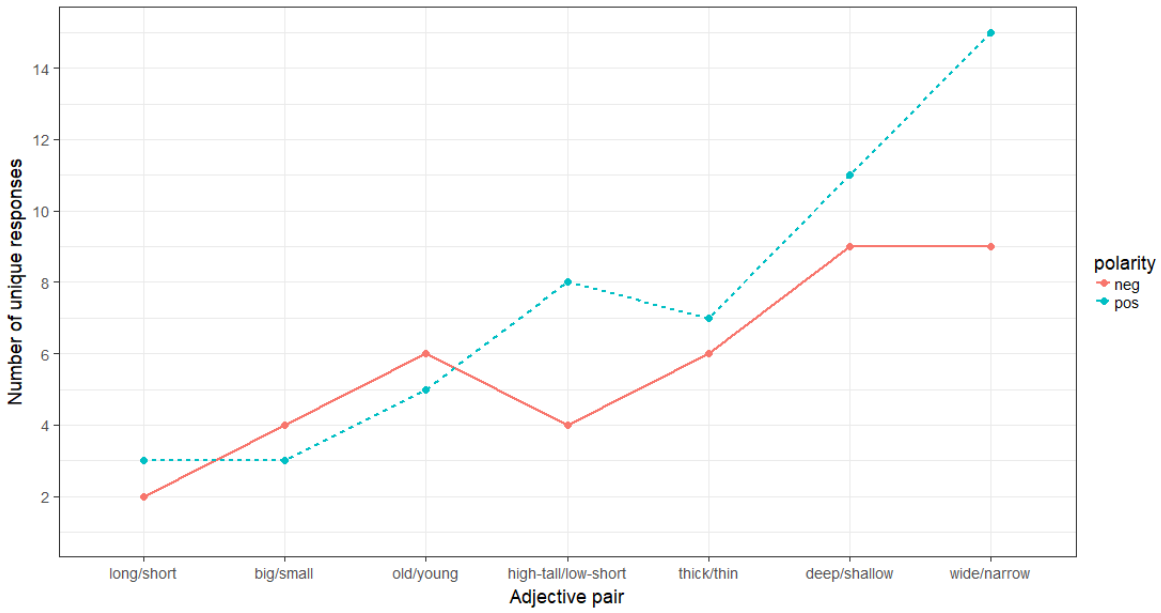


Figure 5.7: Number of unique responses to each negative (solid red line) and positive (dashed blue line) adjective, including the two-year-old's responses. Note that the line for the negative adjectives for the most part remains below the line for the positive adjectives, indicating fewer unique responses.

5.3.3 Discussion

This study examined the production of dimensional antonyms in Hebrew by children between the ages of three and eight. The youngest children who participated in the study are slightly older than the age at which productive adjective use typically begins, (around 1;8, Tribushinina et al., 2014). Therefore, this study captures the process of building up a semantic field of related words starting from its beginning. This study provides a fine-grained picture of the acquisition process, revealing it to be structured in such a way as to reflect the structure of the semantic field.

While the pilot study conducted with three two-year-olds showed that the task chosen was not appropriate for children that young, the data collected provides insight into the initial stages of the construction of the semantic field of dimensional adjectives. Even the youngest children demonstrated knowledge of the contrast between *gadol* ('big') and *katan* ('small'), though only one other AC response was provided by this age group, to *kacar* ('short-l'). The two-year-olds made a large number of errors, which may indicate that they did not yet even know that these words belong to the category of dimensional adjective, though these errors may also be due to various performance issues, or lack of understanding of the task. The responses provided did, minimally, indicate that either the children knew that these words belonged to the category of adjective or that the *which* question required an adjective for an answer.

Starting with the three-year-olds, many more SA responses were provided, though the rate of AC responses remained low. This indicates that while the children do not yet seem to have a fully fleshed out lexical representation of the adjectives, they do know which semantic field they belong to and their polarity, and they can provide a semantically appropriate response to the prompt. The most popular response strategy at this age is to substitute the least specified and, therefore, least complex dimensional adjective for the AC antonym (i.e. *big/small* depending on the polarity of the AC response).

I found that, as children develop, they get better at providing the AC antonym for dimensional adjective prompts. However, their age is not a good predictor of the number of SA responses they provide. By age three, they are already fairly successful at providing a response which contrasts with the adjective in the prompt,

using a number of different and evolving strategies. This result is different from the one found byClark (1972), who reported that both SA and AC responses increased significantly with age. I propose that the reason children in my experiment are better at providing an SA response, despite being younger, is that the task in my study was anchored in a context containing visual, prosodic and syntactic contrast. Such a context, which is similar to the context these adjectives are often found in when used in natural speech, provides an anchor that allows children to figure out what kind of response is needed. Despite not yet necessarily knowing the AC antonym, they are able to figure out enough of the meaning components of dimensional adjectives (see Table 5.1) to provide an appropriate response.

I found initial evidence[11] that the strategies children use to provide an SA response when they did not know the AC antonym change as the children grow older. The most popular strategy for the three-year-olds was substitution with *gadol* ('big') and *katan* ('small') (SB, (19b-ii)). For the four-year-olds, the most popular strategy was to negate the adjective provided (Neg (19b-i)). It is possible that negation is viewed as providing a more precise response than SB substitution. It targets the dimension mentioned, while still not providing the correct antonym, whereas *gadol* and *katan* do not refer to a specific dimension, making them more general and therefore less precise. By age five, children provide the AC response around 70% of the time.

I found no significant differences between the positive and negative members of the antonym pairs. Children provided SA responses much of the time, showing that they are able, from the earliest age in the study, to figure out the correct polarity of the adjective response required much of the time. However, there is suggestive evidence that the negative members of the pairs are more difficult than the positive ones. The youngest child to respond in an AC manner to a negative polarity adjective was, in all but one case, younger than the youngest child to respond correctly to a positive polarity adjective. A response to the marked negative adjective requires the retrieval of the unmarked, more frequent, positive adjective. The negative adjective may also be more strongly associated with the positive adjective than vice versa. It may appear

[11] As discussed in Section 5.3.2, the sample size for each age group in my study is too small to allow definitive conclusions to be drawn. In a larger sample, the trends that I observe could be used to make predictions for the trajectory of developing response strategies.

less frequently without its positive counterpart than the positive adjective appears without the negative one.

The order of acquisition found here for Hebrew matches fairly closely the order predicted by the semantic complexity hypothesis, which is also the order found by Clark (1972) for English. The one exception, as discussed in detail in Section 5.3.2, is *aveh* ('thick') which was predicted to be more complex than *gavoa* ('high/tall'), but I did not find the scores for these two adjectives to be significantly different. Further corpus work is needed to determine whether there are more subtle usage and/or frequency differences between the languages which might explain this discrepancy.

Setting aside this one difference, the overall order of acquisition in the two languages is strikingly similar, despite their typological differences. This similarity provides additional support for the hypothesis that semantic complexity plays a significant role in determining the order of acquisition of these terms. Hebrew differs significantly from English phonologically, morphologically and syntactically in terms of these adjectives and the constructions they enter into; nevertheless, I found the order of acquisition predicted in Clark (1972), based on the theory of semantic complexity proposed by Bierwisch (1967), to hold for Hebrew in almost exactly the same fashion as for English. The substitutions observed also support the complexity hypothesis, with less complex terms (e.g., *gadol/katan* ('big/small')) often substituted for more complex terms, but only once vice versa. In one case, a five-year-old gave the response *namux* ('short-t') to the question *Elmo has a big elephant. What kind of elephant does Cookie Monster have?* In this case, this response was in fact true, and more specific than the expected AC answer. Other than this single exception, none of the dimensional substitutions were more complex than the AC response, as predicted.

There were a number of small differences observed that may be due to the differences between the languages. Children provided more correct answers to *aveh/dak* ('thick/thin') than they did in English, and for the pair *zaken/cair* ('old/young') there was a very large gap between the proportion of SA (.94) and AC (.43) responses in Hebrew, but no such discrepency was found in English by Clark (1972).

Despite these localized differences between the Hebrew and English dimensional adjectives, the overall pattern of the order of acquisition results I found for Hebrew was very similar to that previously established for English (Clark, 1972), providing cross-linguistic support for the hypothesis that semantic complexity, as proposed by Bierwisch (1967), plays a key role in determining the order of acquisition of the highly structured domain of dimensional adjectives.

Overall, the results of this experiment provide a more fine-grained picture of the development of the semantic field of dimensional adjectives. By age two, children can use *gadol* ('big') and *katan* ('small') in contrast with each other, and gradually add in the other dimensional adjectives in order of semantic complexity. Initially the other terms are not even situated in the semantic field, and then as children develop, they start constructing lexical entries for these adjectives, situating them in the correct semantic field and providing the correct polarity even before they have a fully fleshed out, adult-like entry and before they can provide the AC antonym. The visual context provided by the task makes it easier for them to retrieve an SA response, which shares at least some semantic features with the target response, before they are able to provide an AC response. The properties that are easier to acquire can be learned by observing the adjectives used in context. The AC antonym, however, has some degree of lexical idiosyncrasy, and can only be learned by observing how the adjective relates to other adjectives in the semantic field.

5.4 Conclusion

Children start building their mental lexicon before they have fully adult-like lexical entries for words. This study used the highly structured semantic field of dimensional adjectives to examine the formation of one area of the mental lexicon. The gradual construction of the field is exemplified in the shift from making many errors, through substituting the most general dimensional adjectives *gadol* ('big') and *katan* ('small') for more complicated ones and other SA response strategies, to being able to provide the AC antonym.

While children develop adult-like representations of dimensional adjectives gradually, the contextual support provided in this task allowed them to retrieve some of the meaning components of the adjectives. They were able to provide an SA response most of the time as young as age three, long before they can provide an AC response for some of the dimensional adjectives. In previous work, when no context is provided, SA responses were less common among the younger children. The improvement I found in my experiment argues that context provides an important tool to help children understand and produce words they are still in the process of acquiring.

I also found cross-linguistic support for the hypothesis that semantic complexity (Bierwisch, 1967) plays a significant role in driving the order of acquisition of dimensional adjectives as predicted in Clark (1972). I found the order of acquisition in Hebrew to conform for the most part to the order predicted by the complexity hypothesis, similar to previous findings for English (Clark, 1972). Additionally, I found that these order of acquisition effects hold when looking at a wider age range, closer to the age at which children first start using adjectives, around age two. Even the youngest children in my sample, who were too young to complete the task in the experiment, showed evidence of having some representation of the meaning of *gadol* ('big') and *katan* ('small'), the first dimensional adjectives to be acquired. Their errors, however, provide evidence that they do not yet situate the other dimensional adjectives within this semantic domain. Errors and substitutions provided across the age range examined in this study provide evidence of the development of the semantic field of dimensional adjectives as children age; children demonstrate knowledge of the polarity of these adjectives and the fact that they denote dimensionality before they can provide the AC antonym. Properties which can be inferred from the relationship between the adjective and the context in the world it is being used to describe are acquired before properties, such as antonymy, which also require experience with how the adjective relates to other adjectives in the semantic field.

Chapter 6

Discussion

6.1 Introduction

Context plays a significant role in the use and interpretation of relative adjectives. Specifically, the comparison class involved in their interpretation is highly context dependent. The main goal of this dissertation was to explore the notion of a comparison class, examine the interaction of world knowledge and information from the immediate context in its determination, and investigate the components of world knowledge participating in this process. A secondary goal was to examine the acquisition of dimensional adjectives, gaining another perspective on the influence of context. Two studies were presented to further these goals. The first study consisted of a series of experiments investigating adult interpretation of the relative adjectives *big* and *small*. The second study was an acquisition experiment involving dimensional adjective antonym elicitation. The results of the former are the focus of this chapter. I discuss their implications on two different levels: both theoretical implications for the understanding of comparison classes and the semantics of dimensional adjectives, and also methodological implications for the investigation of adjective meaning.

6.2 Summary

In this dissertation, dimensional adjectives are taken as exemplars of the class of relative adjectives. Standard formal accounts of the semantics of relative adjectives propose that they are interpreted relative to a standard of comparison and comparison class – both contextually determined. In order for an object to count as *tall*, for instance, it needs to have a degree of height that exceeds a standard of comparison. This standard is determined by applying some function to the comparison class. Both the standard of comparison and the comparison class are thus crucial to the interpretation of relative adjectives.

Much experimental work has been focused on the standard of comparison. One line of investigation, taken by researchers using different approaches, has been to ask what function is applied to the comparison class to determine the standard of comparison (e.g., Schmidt et al., 2009; Solt and Gotzner, 2012). Regardless of how the standard is set, it is applied to the comparison class, so without knowing what the comparison class is, it is difficult to appropriately interpret the adjective.

Previous literature is in agreement that the comparison class used in the interpretation of relative adjectives is context dependent. However, there is no consensus, and indeed few attempts to reach one, regarding the exact nature of the comparison class. The goal of the experiments I conducted is to explore what kind of contextual information goes into determining the comparison class.

Previous researchers have, at least for experimental purposes, operationalized the comparison class as the immediate visual context, i.e. the items in the display shown to a participant in a given trial (e.g., Hansen and Chemla, 2017; Solt and Gotzner, 2012). However, there is reason to think that the comparison class is not made up exclusively of information from the immediate visual context. Tessler et al. (2017) propose a probabilistic account for how the comparison class of a bare adjective, as in the sentence *It is warm*, is influenced by the context. In this study, the context included information such as the time of year at which the sentence is uttered, and world knowledge regarding temperature distributions. Tribushinina (2011), looking at the adjectives *big* and *small*, provides initial evidence that speakers are integrating

world knowledge with information from the immediate visual context in order to form the comparison class. I based my adjective interpretation experiments on this foundation provided by previous work.

I, like Tribushinina (2011), focused on the adjectives *big* and *small* as representative dimensional adjectives. I adopted the position that speakers integrate world knowledge with information from the immediate visual context when determining a comparison class (the Integration Hypothesis), provided further support for this position and then built on it. With a series of adjective interpretation experiments, I show how different types of world knowledge are utilized depending on the type of stimuli.

In Experiment 1a, I replicate the initial results from Tribushinina (2011) in an experiment conducted on Amazon's Mechanical Turk (MTurk) web-based platform rather than in the controlled environment of a lab and using English rather than Dutch. Participants were shown a series of displays, each consisting of seven images of an object, in a range of sizes, and asked *Which [objects] are [big/small]?*

Two predictions follow from the Integration Hypothesis. The first prediction is that, overall, the mean number of images selected as *big* (the *big* zone) will be smaller than the mean number of images selected as *small* (the *small* zone). The reasoning behind this prediction is that world knowledge regarding the size of objects would dictate that because the images are all smaller than the objects they depict, they should all count as *small* in contrast. However, the visual context presents a bounded set of sizes of the objects, which encourages labeling some of the images as *big* and others as *small*. Integrating these two types of information is argued to cause the asymmetry between the *big* zone and the *small* zone, favoring the *small* zone.

The second prediction relates to speakers' world knowledge about the prototypicality status of objects, i.e. whether they are considered prototypically big objects, prototypically small objects, or objects which are neither prototypically big, nor prototypically small (designated as neutral objects). For example, elephants are considered to exemplify prototypically big objects, while mice are considered to exemplify prototypically small objects. In addition to the overall difference between the *big* zone and the *small* zone, the zone sizes are also predicted to show prototypicality effects,

such that the *big* zone is predicted to be smaller for prototypically big objects than for prototypically small objects. Conversely, the *small* zone is predicted to be smaller for prototypically small objects than for prototypically big objects. Both of the zone sizes for the neutral objects were predicted to fall in between those of the other two prototypicality status categories. These effects provide further support that world knowledge about the objects depicted is taken into account when determining the comparison class.

The results of Experiment 1a support both of these predictions. The *big* zone was found to be smaller than the *small* zone (see Table 6.1). Additionally, prototypicality effects were also observed as predicted. The *big* zone was found to be smaller for prototypically big objects than for prototypically small objects, and the *small* zone was found to be smaller for prototypically small objects than for prototypically big objects.

Experiment 1b demonstrates that these prototypicality effects for the most part persist even when all of the images are actually smaller than the objects they depict. In this experiment, two of the prototypically small objects were replaced with slightly bigger objects that are still considered prototypically small. The replacement is necessary because the largest of the seven images of these two original objects used in Experiment 1a are about the same size as the objects they depicted, in violation of one of the premises relied on in interpreting the results. The difference in zone size between prototypically small and neutral objects was significant when the adjective in the prompt was *small*, though the difference was not significant when the adjective was *big*. As predicted, the *small* zone is still bigger than the *big* zone (Table 6.1).

In Experiments 2 and 3, I demonstrate that these prototypicality effects do not persist when the stimuli are photographs rather than clipart-style images. I speculate that one of the relevant differences between clipart images and photographs is that clipart images do not necessarily depict an actual exemplar of the object category, while photographs, by their very nature, must depict an actual object that existed in the world. Clipart images may instead depict a generalized, idealized, typical object. Because of this difference, it may be that clipart stimulus images trigger access to a comparison class made up of representations of average, prototypical members of the

category, which do not actually exist, and are not associated with actual dimensions. Rather, the representations in this comparison class are generalized members of the class. Therefore, the distribution of sizes of these representations is rather narrow and restricted.

Photographs, on the other hand, must be representations of actual objects which were associated with actual dimensions. They may trigger representations in the comparison class including actual objects of the depicted category which the speaker may have interacted with, and which also had actual dimensions. This distribution of dimensions may contain more outliers and be more widely distributed around the mean.

This difference in the composition of the comparison class may cause the different patterns observed in Experiments 1a and 1b vs. Experiments 2 and 3. In Experiments 1a and 1b, there is a distinction between prototypically small and neutral objects, while in Experiments 2 and 3 there is no significant difference between these two categories. It is possible that the distinction found in Experiments 1a and 1b, when the stimuli were clipart images, is due to the comparison class distributions for prototypically small objects and neutral objects each being fairly narrow, and therefore they overlap with each other less. The lack of distinction between prototypically small and neutral objects when the stimuli are photographs may be due to the wider distribution of sizes in the comparison class, and therefore greater overlap between the categories. The prototypically big objects are bigger in the real world than the small or neutral objects, to such an extent that even with a wider distribution reflecting a realistic array of actual sizes there is little overlap between the size of objects in this category and those in the other two categories. Despite the lack of significant prototypicality effects when the stimuli are photos, overall the *small* zone is bigger than the *big* zone, as predicted following the same reasoning presented for Experiment 1a (Table 6.1).

An alternate explanation is considered for the difference between the size of the *small* zone and the size of the *big* zone in Experiments 1–3. The *small* zone might have been bigger than the *big* zone because the scale structures of *big* and *small* are not identical (the Scale Structure Hypothesis). However, Experiments 4 and 5 provide further support that the effects observed in Experiments 1–3 are not simply

due to scale structure differences, but rather, as originally argued, they are related to the integration of world knowledge of some form with the information from the immediate visual context (the Integration Hypothesis). The stimuli for these two experiments were images that are larger than the objects they depict, such as an ant or a blueberry. In this case, world knowledge indicates that all of these images should count as big, while the visual context contributes the same information as in the previous experiments. The integration of world knowledge with information from the immediate visual context predicts the reversal of the asymmetry between the *big* and *small* zones observed in these experiments, whereby the *big* zone is the larger of the two, as shown in Table 6.1.

Exp.	Image type	Order	Relative size of object to image	*Big* zone	*Small* zone	*Big* zone > *small* zone?
1a	clipart	ordered	object bigger	2.11	2.87	N
1b	clipart	ordered	object bigger	2.21	3.03	N
2	photograph	ordered	object bigger	2.33	3.14	N
3	photograph	random	object bigger	2.49	2.89	N
4	photograph	ordered	object **smaller**	3.81	1.54	**Y**
5	photograph	random	object **smaller**	3.49	1.96	**Y**

Table 6.1: Comparison of the mean *big* and *small* zones in Experiments 1–5

The reversal of the zone sizes can be seen clearly in the results summarized in Table 6.1. In the bottom two rows, in the experiments in which the images were larger than the objects depicted in them, the *big* zone is bigger than the *small* zone. In the first four experiments, the reverse is true. The *big* zone is smaller than the *small* zone. This reversal in zone sizes supports the Integration Hypothesis whereby information from the immediate visual context is integrated with world knowledge to determine the comparison class used in relative adjective interpretation.

In order to accomplish the secondary goal of this dissertation, I investigated the acquisition of dimensional adjectives as a semantic field. As children build up their lexicon, there are a number of meaning components that they must learn for each of

the dimensional adjectives, including that they relate to dimensionality in general, the specific dimension they denote, their polarity and the appropriate associated antonym. Forming this association requires understanding the contribution of all of the adjectives' meaning components. All of this information provides structure to the semantic field. I show that the structured nature of this field is reflected in the acquisition process. Here too, context plays an important role. Children start using dimensional adjectives around age two, and from very early on in the acquisition process they are able to use members of antonym pairs in contrast with each other. However, previous work has been unsuccessful at eliciting these pairs before age four. In my study, the stimuli were scenes containing object pairs that contrast with each other only in the property named by the adjective in the prompt. The visual, prosodic and syntactic contrast in the stimuli allows children as young as age three to access their lexicons and retrieve the antonym. Without the picture, three-year-old children have been found to be unable to retrieve the associated antonyms (Clark, 1972).

Additionally, even with the pictures, especially for the semantically more complex adjectives, some of the younger children provided responses that were not even dimensional adjectives, but rather color or shape adjectives. One possible explanation is that they may not even have associated the adjective in the prompt with the dimensional domain. At the very least, it indicates that the association between the adjective in the prompt and the dimensional domain was not strong enough to trigger an association with another dimensional adjective. Moreover, the fact that the older children rarely made this type of mistake suggests that they are attending to the word uttered and not just to the picture displayed.

I argue that the various meaning components of dimensional adjectives are not acquired all at once, as complete lexical entries. Rather, the fact that semantically appropriate responses are provided before adult correct ones suggests that the polarity and dimensionality of these adjectives is acquired before the complete association of the adjective with its antonym. I also found the order of acquisition of these adjectives to be consistent with the complexity hierarchy proposed by Clark (1972). This result provides support for the hypothesis that semantic complexity is a crucial factor in determining the order of acquisition of dimensional adjectives.

6.3 Theoretical Implications

This section explores three areas in which my adjective interpretation experiments have theoretical implications, framed as a discussion of the generalizability of my results to concentric circles of influence: other stimuli, other adjectives, and other linguistic domains.

The first question I address is whether my results generalize to other stimuli using different objects and images. My work starts to explore the relation between world knowledge relating to prototypicality and knowledge of properties such as real-world size, accumulated through experience with objects in the world. However, further work remains to be done to further distinguish between the different types of world knowledge utilized in the interpretation of relative adjectives and how they are weighted. In Section 6.3.1, I discuss some of the open questions in this area by addressing the generalizability of my results across different types of stimuli.

The second question I address is whether my results generalize to broader classes of adjectives. In this dissertation, I have taken dimensional adjectives to be exemplars of the class of relative adjectives. In Section 6.3.2, I discuss whether this move is justified, and whether my results might have implications even beyond this class, to other classes of gradable adjectives as well.

Finally, I address the question of whether my results generalize beyond adjective interpretation to broader domains. Based on the series of adjective interpretation experiments I conducted, I argued in support of the Integration Hypothesis: world knowledge is integrated with information from the immediate visual context when determining a comparison class. I provided evidence that different types of world knowledge are used, and argued that the influence of different types of world knowledge on the comparison class can vary depending on features of the visual context. These results touch on broader questions of how information from multiple sources is integrated to form one input to linguistic processes. Other work has approached similar questions from very different angles. In Section 6.3.3, I explore one such angle and discusses how my work fits into this broader question, and perhaps generalizes to other domains.

6.3.1 Generalizability across Objects and Images

Experiments 1–3 used a specific set of stimuli which were chosen, for the most part, by Tribushinina (2011). These stimuli are taken to be examples of three prototypicality status categories (prototypically big objects, prototypically small objects, and objects which are neither prototypically big nor prototypically small). The selection process described in Tribushinina (2013a) as well as my norming study (Chapter 2), were intended to ensure that indeed these objects had the prototypicality status attributed to them.

Not only were specific objects chosen as stimuli, but specific images were chosen to represent these objects, both clipart-style images in Experiments 1a and 1b, and photographs in Experiments 2 and 3. The interpretation of my results depends on the images chosen being representative of the object categories and the objects chosen being representative of the prototypicality status categories. The results are predicted to hold up regardless of the specific images and objects chosen if the images are in fact all representative of their respective categories.

The issue of generalizability across images is particularly relevant when discussing the distinction between prototypicality status and real-world size. I have proposed that clipart-style images are more stereotypical, as they do not tend to depict an actual object, but rather an average object of the category. The images used as clipart stimuli in Experiments 1a and 1b were not consistent as to how realistic the depiction of the object is. Some, like the elephant, were fairly accurate representations of the object. Others, like the mouse, which was an image of a silhouette of a mouse with no details at all, were less accurate. Photographs, on the other hand, must depict actual objects which are associated with actual sizes.

Many of the objects selected as stimuli exist in a fairly large range of sizes. For example, a plane could be a tiny one-person airplane or a jumbo jet seating hundreds. A house might range from a small cottage to a ten-bedroom mansion. This range might be utilized to provide further insight into the distinction between prototypicality status and real-world size information that is used to determine the comparison class. My results thus far predict that the clipart conditions should be unaffected by changing the specific images used. In other words, if clipart images really are

generalized, stereotypical depictions, and less strongly related to the dimensions of an actual object, then using a different image might still be expected to access the same comparison class representation. Thus, the prototypicality effects observed in Experiments 1a and 1b should be observable regardless of which clipart images are used for the objects. However, photographs, which depict actual exemplars of the category, might more easily trigger real-world size information related to the dimensions of the actual exemplar in the image. The photograph condition might then be more affected by changing the specific images used. If other images of the objects are used, the results might be expected to change significantly.

The issue of generalizability across objects is related to a larger question regarding the nature of prototypicality status. Prototypicality status is proposed to play a role in determining the comparison class of relative adjectives. It seems to be part of the world knowledge speakers possess about objects in the world. However, the experiments in this dissertation were conducted on a fairly restricted set of objects. I showed that elephants, houses, planes, and hippos as a group are different from bunnies, ducklings, babies and gnomes with regard to zone size, and therefore comparison class. I argued that the relevant distinction between these two groups lies in their prototypicality status, which predicts that this pattern should generalize beyond these specific objects. Other objects similarly categorized as prototypically big or prototypically small objects are predicted to give rise to the same prototypicality effects observed in Experiments 1a and 1b. Conducting similar experiments using different objects as stimuli would provide further support for the claim that prototypicality status is relevant to comparison class determination. It would also provide further insight into the nature of prototypicality status as a cognitively salient and relevant concept.

An interesting direction to follow up on would be to examine the consistency of these experiments' results when varying the images of the objects, as well as the objects themselves. Varying the images may provide further insight into the distinction between real-world size and prototypicality status proposed in Chapter 3. Varying the objects may provide further insight into prototypicality status as a relevant factor in determining the comparison class.

6.3.2 Generalizability across Adjectives

In the work presented in this dissertation, dimensional adjectives were taken as exemplars of relative adjectives in general. This choice raises two major questions. The first question is whether it is actually justified to use dimensional adjectives as exemplars of relative adjectives more generally. This question will be addressed in the first part of this section, which discusses the relationship between dimensional adjectives and other relative adjectives, and examines whether the differences between the different types of relative adjectives are relevant to my findings. The second question is whether my findings have any applicability outside the domain of relative adjectives. It will be addressed in the second part of this section, which discusses whether the work in this dissertation is at all relevant to the interpretation of absolute adjectives, which, like relative adjectives, are gradable, but are less (or not at all) context dependent (see Section 1.2 for background on this distinction).

Generalization to Relative Adjectives

Dimensional adjectives describe properties that are concrete and easy to represent visually, and it is easy to order objects according to the degree of the property they possess. Therefore, throughout this dissertation, these adjectives have been used as a proxy for the more general class of relative adjectives. In order for this move to be justified, dimensional adjectives must share all of the relevant properties with other relative adjectives.

The semantic account of dimensional adjectives is not typically distinguished from that of other relative adjectives (see Section 1.2). Like other relative adjectives, the interpretation of dimensional adjectives is dependent on a standard of comparison, applied to a contextually determined comparison class.

However, there may be one difference between dimensional adjectives and other relative adjectives. Relative adjectives in general are taken to have scales which are open at both ends (Kennedy, 2007; Kennedy and McNally, 2005), i.e. do not have endpoints. For example, the scale of beauty, with *ugly* at one end and *beautiful* at the other, is doubly open. There is no limit to the degree of beauty that one may

possess, in either direction (see Section 4.1). However, it is less obvious that dimensional scales such as size share this property. There seems to be a clear zero point. It has been proposed that all scales of relative adjectives, including dimensional scales such as size, are asymptotically bounded at zero on the lower end (e.g., Kennedy and McNally, 2005; Lehrer, 1985). Therefore, dimensional adjectives should behave similarly to other relative adjectives. But others have distinguished between relative adjectives that are asymptotically bounded at zero and ones that are truly unbounded, going to negative infinity (Klecha, 2012; Lassiter, 2017). If such a distinction is made, the question arises whether unbounded adjectives behave the same as asymptotically bounded dimensional adjectives in terms of the affect of context on their interpretation.

The notions of a contextually determined standard of comparison and comparison class are still crucial to the interpretation of all relative adjectives. Since my work focuses on the comparison class, it is plausible that this scale structure distinction would not affect the arguments I present in this dissertation. Therefore, my findings should generalize to other relative adjectives, including those with no lower bound. Further work is needed to demonstrate that this is indeed the case. One direction this work could take is replicating Experiments 1–5 with non-dimensional relative adjectives, which would face methodological challenges with stimuli design, but may be possible.

Generalization beyond Relative Adjectives

Relative adjectives are also part of the larger class of gradable adjectives. The second question my work raises is to what extent is it possible to generalize the claims presented here to this broader class of adjectives, or more specifically, to the other major class of gradable adjectives, absolute adjectives. It is a matter of debate in the literature to what extent absolute adjectives, such as *full*, *closed* and *dirty*, are context dependent. Kennedy and McNally (2005) for example, argue that absolute adjectives are not context dependent at all (see also Kennedy, 2007). While relative adjectives have a contextually determined standard of comparison somewhere in the middle of

the scale associated with the adjective, absolute adjectives have an endpoint standard of comparison regardless of the context of use. An opposing position is taken, for example, by McNally (2011). She argues that while the standard of comparison for relative adjectives is determined by classification by similarity, the standard for absolute adjectives is determined by classification by rule. Thus, the difference between absolute and relative adjectives is not context dependence, but rather how the standard of comparison is determined. The ways my work might generalize to absolute adjectives depends on how their semantics are accounted for.

According to the approach that absolute adjectives are not dependent on context at all, there seems to be no reason to expect my findings to be particularly relevant to the interpretation of absolute adjectives. My work focuses on the comparison class, while the semantics of absolute adjectives does not make reference to one. However, while they are not context dependent in the same way that relative adjectives are, even proponents of the position that they are not at all dependent on context agree that they are imprecise. The standard of comparison is not necessarily identical in all cases. For example, when someone utters the sentence in (24), the interpretation might be different depending on whether the context is an operating room, a kitchen or a woodworking shop.

(24) The knife is clean.

The seemingly different standard in these cases has been argued to result from a pragmatic halo (Lasersohn, 1999), regulating how close to the standard is close enough to count as true in each context (Kennedy, 2007).[1] The pragmatic halo is a set of objects that only differ from the precise denotation in ways which are pragmatically ignorable in the context. Thus, if (24) is uttered in an operating room, the halo will be tighter and encompass fewer similar objects, or perhaps contain only an absolutely sterile knife. In a woodworking shop, the halo might be looser, and allow for microscopic contamination, and perhaps even a little bit of visible dirt, and yet the knife might still count as *clean* in this context.

In determining what counts as pragmatically ignorable, it is possible that various

[1]See Kennedy (2007, p. 25, footnote 22) for a summary of other approaches that have been taken to solve the same issue.

factors are taken into account, and a representation similar to a comparison class might be assessed. In the example in (24), the knife might be compared to representations of knives with varying levels of dirt previously experienced. This comparison will influence the size of the halo, in order to group the knife representations into equivalency classes, and then determine which ones are similar enough to be equivalent in the context of the utterance.

Determining the appropriate level of imprecision might be a place where my findings are relevant even to absolute adjectives. On the one hand, it is not implausible that the world knowledge necessary to make these determinations (e.g., knowledge that an operating room should be sterile) interacts with the knives in the immediate context to determine the interpretation of the adjective. On the other hand, it is also possible that, for absolute adjectives, only world knowledge is relevant to setting the size of the halo and the items in the immediate visual context are irrelevant. If this is the case, then my results might have less relevance for absolute adjectives in terms of integration of information from the visual context with world knowledge. Nevertheless, world knowledge of different types might be involved in the interpretation of absolute adjectives, and experiments similar to mine might be helpful in exposing these different types and distinguishing their various influences from each other.

Alternatively, there are accounts of absolute adjective semantics (e.g., Lassiter and Goodman, 2013) which do involve a comparison class. Instead of the comparison class containing representations of different entities of the same object category with differing degrees of the relevant property, it contains representations of different potential degrees of the property in the same object. For example, the knife in (24) might have different potential levels of cleanliness, and those might make up the comparison class against which the standard is determined (Toledo and Sassoon, 2011). The function for setting the standard for absolute adjectives differs from relative adjectives. It is possibly set to the relevant endpoint of the scale (Kennedy and McNally, 2005).

Under the comparison class account of the semantics of absolute adjectives, there may potentially be an interaction between the immediate visual context and world knowledge in determining which are the relevant representations in the comparison class. Perhaps, for instance, in the case of the knife in the operating room, there are

different representations for a completely sterile knife and a knife with only microscopic contamination, whereas in the comparison class for the knife in the context of the woodworking shop there will only be one representation. Assuming an endpoint standard for an adjective like *clean*, the most extreme representation in the comparison class, which in the latter case might not be completely clean, would count as *clean* while the other representations would not.

If such a comparison class is involved in absolute adjective semantics, then it is clear that my results may potentially have implications for the interpretation of absolute adjectives as well. Moreover, even if there is no comparison class involved, there are potential implications for the resolution of imprecision of absolute adjectives.

6.3.3 Generalizability to Other Domains

Beyond the generalizability to other images, objects and adjectives, we might also consider the relevance of this work to broader linguistic domains. Experiments 1–5 support an account in which the information from the visual context is integrated with world knowledge to determine the comparison class used for the interpretation of relative adjectives. Comparison class determination is not the only domain in which integration of information from multiple sources occurs in the interpretation of linguistic stimuli. This section examines one such domain – color word interpretation,[2] and discusses how my proposal for adjective interpretation relates to this work.

Cibelli et al. (2016) propose a probabilistic model of cue integration for the color domain. After describing their model, I argue that the cue integration that they are modeling can be thought of as parallel to the integration of information from various sources observed in Experiments 1–5. These parallels help integrate my proposal into a much broader literature, and may provide interesting further directions to my work. The model proposed by Cibelli et al. (2016) could be used to make more precise predictions for further experimentation regarding the integration of different sources

[2]While color words are expressed as adjectives, as will be described in detail, this proposal relates to the more perceptual aspect, and is built on work even farther afield, from integration of visual with haptic cues, as well as work on category effects in various domains, including spatial location, object size and vowel perception.

of information when determining the comparison class.

Cibelli et al. (2016) model color perception and color recall using a category adjustment model (Bae et al., 2015; Persaud and Hemmer, 2014). Color recall involves the probabilistic combination of information from two sources: a fine-grained representation of the particular color seen, and the stored representation of the range of the spectrum covered by the linguistic category label of the color seen (e.g., *green*). Both the fine-grained representation of the color stimulus seen and the grosser representation of the color label are modeled as probability distributions over a universal perceptual color space. The fine-grained representation of the stimulus is centered at the exact location of the color seen, while the variance of the distribution represents the observer's uncertainty about where that exact location is. In cases of greater uncertainty, the variance is greater as well. The uncertainty may be due, among other things, to noise in the perception itself (e.g., because of blurry vision, bad light), or to memory decay over time, all resulting in a wider, flatter distribution.

The linguistic color label is also modeled by a probability distribution over the universal color space. It covers the entire space included under the color label, with higher probability given to more prototypical instances of the label. The original location of the stimulus is inferred by combining the information from both of these distributions, which tends to produce reconstructions that are biased away from the actual location, and toward the prototype of the color label in which the stimulus falls. This bias results in more differentiation for the stimuli across category boundaries, than for stimuli that fall under the same color label, similar to category effects in memory (Bae et al., 2015). Take, for example, two stimuli, one which would be labeled *green* and another which would be labeled *blue*, that are the same distance apart from each other in the universal color space as two other stimuli, both labeled *green*. The two with different labels will be recalled as farther apart from each other in the universal color space than the two stimuli which share a label. This account is framed in terms similar to my proposal, albeit in a different domain, and therefore it is worth exploring the parallels.

Cibelli et al. (2016) discusses the integration of two cues from the visual domain – the color perceived, as situated in memory on the color spectrum, and the linguistic

category label the stimulus falls under, e.g., *green*. Experiments 1–5 can also be conceptualized in terms of the integration of multiple cues. Just as a universal color space was posited in Cibelli et al. (2016), which is a mental representation of the actual spectrum of colors, I propose there might be a universal size scale. The universal size scale would be a mental representation of actual size, covering the entire size possibility space. In my experiments there are two things happening simultaneously. First, participants need to figure out how the stimuli in the context correspond to actual size. Additionally, given the stimuli and world knowledge about the objects seen, they need to determine a comparison class and standard of comparison in order to determine what counts as *big* and *small*. The way the stimuli correspond to real-world size influences the determination of the comparison class and the standard.

In order to determine the correspondence to real-world size, and then the comparison class and the standard of comparison, there are two cues that need to be integrated, as there were in Cibelli et al. (2016). One cue is the visual information in these experiments, the set of images which need to be situated somewhere on the mental size scale along with the rest of the mental representation of the distribution of items in that category. This cue is comparable to the color perceived. Both the color seen and the visual information from the display are objective facts, about which judgments are made.

The second cue, unlike the second cue in Cibelli et al., is not visual. Rather, it is the world knowledge regarding the size of the objects in the world, which can be compared to the color label cue from Cibelli et al. (2016). Despite the differences, there is similarity between the two in that this cue varies across speakers and is more subjective. The variation discussed in Cibelli et al. was due to different languages dividing up the color space differently and having different boundaries for color categories. In my experiments the variation is introduced since there are different types of world knowledge which do not result in the same comparison class, and therefore do not always line up in terms of where the standard of comparison should be placed in the comparison class.

Uncertainty is introduced by the need to link the images observed with the mental

scale of size. It is not always straightforward to map an observed image to a real-world size. Varying levels of uncertainty might result from different types of stimulus images. I propose, based on the differences in the results between Experiments 1a–b and 2, that when the images are clipart there may be more uncertainty as to how the size of the object depicted maps to the size scale, since the images do not necessarily depict a specific object that actually exists in the world. When the images are photographs there may be less uncertainty, since the images depict objects that actually existed, and therefore necessarily had actual dimensions, and there may be additional information in the photograph that helps retrieve those actual dimensions. When the size of the image is close to the actual size of the object, as is the case with the tiny things, there may be even less uncertainty as to how to map these images to real-world sizes. It is more plausible to simply map the size of the image to the size of the object depicted. The lower uncertainty may account for the stronger effect found in Experiments 4 and 5.

As the uncertainty over the size of the images and how they fit into the distribution of sizes of objects in the comparison class representation increases, the bias to judge the size of a stimulus as closer to the size of the prototype increases, and therefore prototypicality effects, similar to the category effects described in Cibelli et al. (2016) emerge. When there is more certainty, such as when the images are photographs, it is clearer how to map the images viewed onto the distribution making up the comparison class. The categories are less clear-cut; instead, there is more of a continuum mapping to actual sizes because the possible distribution of sizes of actual objects of that category is wider, with more outliers. Averaged across object categories the distributions are still wider and flatter, and therefore the zone sizes for prototypically small and neutral objects are not distinguishable, since their real-world sizes are fairly similar. The zone sizes for small and neutral objects are still distinct from the zones of prototypically big objects which have very different actual sizes.

There are two advantages provided by this parallel between Experiments 1–5 and the work in Cibelli et al. (2016). The first advantage is the link provided between my work and a broader literature on cue integration. If my work can be integrated into this framework, and all the phenomena in both areas accounted for by a single

model, this provides a more economical theoretical landscape. A more general account of processes of language interpretation, relevant to more than one subdomain, is beneficial in general for reasons of theoretical economy. The second advantage is that the work in Cibelli et al. (2016) and others operating in this framework may provide a more concrete basis for generating hypotheses and predictions for future work. The model outlined in Cibelli et al. (2016) may provide a foundation for examining comparison classes in terms of uncertainty, and may help generate predictions regarding the behavior of comparison classes in conditions of increased and decreased certainty. These in turn may be used to support more specific accounts of the integration of different sources of information when determining the comparison class.

6.4 Methodological Implications

In addition to the theoretical implications, Experiments 1–5 also have methodological implications for experimental work on adjective interpretation. Two main paradigms that have been previously used but should be reconsidered in light of my findings are the visual world paradigm (e.g., Tanenhaus et al., 1995; Sedivy et al., 1999) and the paradigms used in studies of standard setting and scale structure, such as Solt and Gotzner (2012); Barner and Snedeker (2008); Hansen and Chemla (2017) (see Section 1.4.1). Much of the empirical work on adjective semantics relies heavily on the images in the display shown to participants as the main cue to the comparison class. While this method has the benefit of making experiments easier to design and interpret with regard to the standard of comparison, it also poses a limitation because it oversimplifies the picture. As I have argued in this dissertation, adjective interpretation is more complex, and the multiple sources of information interact with each other to determine the comparison class. This section considers previous experimental work on adjectives in light of my findings.

6.4.1 Implications for Visual World Experiments

Work using visual world paradigms, such as Sedivy et al. (1999), which examines the effect of adjectives on disambiguation during processing, can be reinterpreted in light of the work presented in this dissertation. In these visual world paradigms, the typical experimental display for a size adjective consists of a square with an object in each corner. Two corners contain an object identical but for its size, the target and the contrast objects. One corner contains an object which competes with the target as a referent of the adjective modified noun phrase (i.e. to which the adjective can also felicitously be applied), and the final corner contains a distractor object, as in Figure 6.1 (Sedivy et al., 1999, p. 130, Fig. 4). In the experiment in Sedivy et al. (1999) there is also a non-contrasting condition in which the contrast object is replaced by an additional distractor (e.g., a file folder). Participants heard a sentence such as *pick up the tall glass and put it below the pitcher*, and their eye-movements throughout the utterance were recorded.

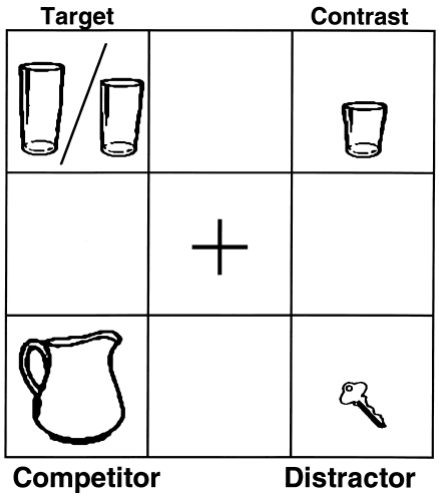


Figure 6.1: Sample visual world display, for the sentence 'pick up the tall glass and put it below the pitcher', from Sedivy et al. (1999, p. 130, Fig. 4)

This experiment demonstrates that information regarding the contrast in the display (e.g., the presence of a tall glass and a short glass) is used in conjunction with the adjective (*tall*) to identify the correct referent before the onset of the noun (*glass*).

Participants looked at the target more than the contrast before the processing of the head noun. Additionally, there were more looks at the competitor in displays without contrast than in displays with it.

However, the relative size of the images in these displays is not consistent. For example, in the display in Figure 6.1, the key, which is a distractor, is about two-thirds the height of the tall cup. Thus it is a fairly tall key. Therefore, it should perhaps also be analyzed as a competitor, rather than a distractor. With this specific example, the problem is perhaps less egregious, since *tall* is not the most natural way to describe a key, but the point still stands. The paper does not provide details about the other stimuli (the other adjectives used or the other possible images), or give a sense of their relative sizes.

Sedivy et al.'s main point regarding the contribution of the contrastive interpretation of the adjective most likely still stands in this case, but the baselines should probably be readjusted taking into consideration other potential labels given the relative size of the objects depicted. Additionally, this paradigm has been used frequently to expand on the work in Sedivy et al. (1999) and in some of those follow-up experiments the relative size of the objects may also be relevant. For example, in both Dix et al. (2018) and Kim and Reksnes (2018) a very similar visual world paradigm is used to examine individual differences in using context to resolve reference. In these cases, where the distinctions are subtler, it may be even more important to be clear about which adjectives are applicable to which objects in the display and how the experiment is affected by the relative size of the images.

6.4.2 Implications for Standard of Comparison Experiments

Earlier, I discussed a number of experiments relating to the standard of comparison for relative adjectives. They were presented as part of the motivation for conducting my adjective interpretation studies. This earlier discussion was focused on how the comparison class was operationalized in these experiments and what implications this decision might have on the interpretation of the results (see Section 1.4.1). In light of my results, I revisit these earlier experiments in this section.

In experiments such as the one presented in Solt and Gotzner (2012), the comparison class is clearly operationalized as containing only the objects depicted in the display shown to participants. My adjective interpretation studies provided evidence supporting the position that when constructing a comparison class for the interpretation of a relative adjective, the information in the immediate visual context is integrated with world knowledge regarding the size of the objects in the context. In light of this evidence, in order to investigate the standard of comparison, it is important to carefully control the comparison class. One way to do this is to restrict the stimuli to novel objects, ensuring that participants have minimal prior knowledge about them. Thus, the only information that they have is what is presented to them over the course of the experiment.

Additionally, participants might update this knowledge over the course of the experiment. If this is the case, each trial adds information as to the potential distribution of sizes of objects in the comparison class. As discussed earlier regarding the experiment in Hansen and Chemla (2017), their anomalous result, whereby *tall* did not behave as predicted for relative adjectives, might be accounted for by looking at the other stimuli in the experiment. By the time the target sentences were heard, participants had minimally seen six example displays, all with the aliens of the same height. If these aliens were taken into account in forming the comparison class, then it is plausible that participants found it hard to call aliens who were shorter than the ones in those initial trials *tall*. If the comparison class is thus updated over the course of an experiment, then in addition to controlling for prior information about the stimulus objects that subjects might have, the information presented over the course of an experiment must also be taken into account (see discussion in Section 4.1).

The results of Experiments 1–5 provide reason to exercise caution in selecting stimuli for future experiments involving the interpretation of relative adjectives. These considerations are equally relevant when investigating the comparison class and the standard of comparison.

6.5 Conclusion

In this dissertation, I explored the use and acquisition of relative adjectives, focusing on dimensional adjectives. My main goal was to explore the notion of the comparison class, examine the interaction of information from the immediate context and world knowledge in its determination, and investigate the components of world knowledge. In order to accomplish this I conducted a series of adjective interpretation experiments which, I argued, provide evidence that comparison class representations integrate information from the immediate visual context in which the adjective is used with world knowledge that the interpreter has regarding the objects in the context. I proposed that different types of world knowledge are relevant to the determination of the comparison class, and examined two distinct types of world knowledge. The first type I examined is knowledge of the prototypicality status of objects, namely, whether an object is considered prototypically big, prototypically small, or neither. The second type of world knowledge I examined is knowledge of the actual size of objects in the real world. The results of these experiments support an account in which the type of depiction influences how these two types of knowledge are weighed against each other in the determination of the comparison class.

My secondary goal was to examine the acquisition of the semantic field of dimensional adjectives. Using an antonym elicitation experiment, I demonstrated that contrastive contexts help anchor children's production of dimensional adjective antonyms. Once there is an appropriate context present, children as young as age three are capable of performing the task, providing a more complete picture of the acquisition process of dimensional adjectives. I provided a fine-grained picture of the gradual process of acquiring the semantic field. The experiment showed that children are capable of providing semantically appropriate responses from early on. However, the adult antonym associations are learned later, and children's adult-like antonym production does improve with age. I argue that the various meaning components of dimensional adjectives are not acquired all at once, as complete lexical entries. Rather, the finding that semantically appropriate responses are provided before adult correct ones suggests that the polarity and dimensionality of these adjectives is acquired before

the complete association of the adjective with its antonym.

Together these studies provide a deeper understanding of some of the ways in which various contextual factors interact and influence the use and acquisition of relative adjectives.

Bibliography

Au, Terry K.-f., and Ellen M. Markman. 1987. Acquiring word meanings via linguistic contrast. *Cognitive Development* 2:217–236.

Bae, Gi-Yeul, Maria Olkkonen, Sarah R. Allred, and Jonathan I. Flombaum. 2015. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General* 144:744–763.

Bale, Alan. 2011. Scales and comparison classes. *Natural Language Semantics* 19:169–190.

Barner, David, Neon Brooks, and Alan Bale. 2011. Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition* 118:84–93.

Barner, David, and Jesse Snedeker. 2008. Compositionality and statistics in adjective acquisition: 4-year-olds interpret *tall* and *short* based on the size distributions of novel noun referents. *Child Development* 79:594–608.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68:255–278.

Barsalou, Lawrence W. 1983. Ad hoc categories. *Memory & Cognition* 11:211–227.

Barsalou, Lawrence W. 1987. The instability of graded structure: Implications for the

nature of concepts. In *Concepts and Conceptual Development: Ecological and intellectual factors in categorization*, ed. U. Neisser, 101–140. Cambridge: Cambridge University Press.

Bartlett, Elsa Jaffe. 1976. Sizing things up: The acquisition of the meaning of dimensional adjectives. *Journal of Child Language* 3:205–219.

Bartsch, Renate, and Theo Vennemann. 1972. *Semantic Structures: A study in the relation between semantics and syntax*. Frankfurt a.M.: Athenäum.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1–48.

Berman, Ruth A. 1985. The acquisition of Hebrew. In *The Crosslinguistic Study of Language Acquisition: Theoretical issues*, ed. D. I. Slobin, volume 1, 255–371. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Biederman, Irving. 1987. Recognition-by-components: A theory of human image understanding. *Psychological Review* 94:115–147.

Bierwisch, Manfred. 1967. Some semantic universals of German adjectivals. *Foundations of Language* 3:1–36.

Bierwisch, Manfred. 1989. The semantics of gradation. In *Dimensional Adjectives*, ed. M. Bierwisch and E. Lang, 71–262. Berlin: Springer.

Bol, Gerard W. 1995. Implicational scaling in child language acquisition: The order of production of Dutch verb constructions. In *Papers from the Dutch-German Colloquium on Language Acquisition, Amsterdam Series in Child Language Development*, ed. M. Verrips and F. Wijnen, volume 3, 1–13. Amsterdam: Institute for General Linguistics.

Bolinger, Dwight. 1967. Adjectives in English: Attribution and predication. *Lingua* 18:1–34.

Brewer, William F., and J. Brandon Stone. 1975. Acquisition of spatial antonym pairs. *Journal of Experimental Child Psychology* 19:299–307.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6:3–5.

Burnett, Heather. 2014. A delineation solution to the puzzles of absolute adjectives. *Linguistics and Philosophy* 37:1–39.

Casasanto, Daniel, and Gary Lupyan. 2015. All concepts are ad hoc concepts. In *The Conceptual Mind: New directions in the study of concepts*, ed. E. Margolis and S. Laurence, 543–566. Cambridge, MA: MIT Press.

Christensen, Rune H. B. 2015. ordinal—regression models for ordinal data. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/.

Cibelli, Emily, Yang Xu, Joseph L. Austerweil, Thomas L. Griffiths, and Terry Regier. 2016. The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PloS ONE* 11:e0158725.

Clark, Eve V. 1971. On the acquisition of the meaning of *before* and *after*. *Journal of Verbal Learning and Verbal Behavior* 10:266–275.

Clark, Eve V. 1972. On the child's acquisition of antonyms in two semantic fields. *Journal of Verbal Learning and Verbal Behavior* 11:750–758.

Clark, Eve V. 1973. Non-linguistic strategies and the acquisition of word meanings. *Cognition* 2:161–182.

Cresswell, Maxwell J. 1976. The semantics of degree. In *Montague Grammar*, ed. B. Partee, 261–292. New York: Academic Press.

Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Dirven, René, and John R. Taylor. 1988. The conceptualisation of vertical space in English: The case of *tall*. In *Topics in Cognitive Linguistics*, ed. B. Rudzka-Ostyn, 379–402. Amsterdam & Philadelphia: John Benjamins.

Dix, Sadie, Rebecca Lawrence, Cameron Morgan, and Chigusa Kurumada. 2018. Integration of top-down and bottom-up information in online interpretations of scalar adjectives. Presented at the 92^{nd} Annual Meeting of the Linguistic Society of America, Salt Lake City, UT.

Dixon, Robert M. W. 1982. Where have all the adjectives gone? In *Where Have All the Adjectives Gone?: And other essays in semantics and syntax*, 1–62. Berlin: Walter de Gruyter.

Donaldson, Margaret, and Roger J. Wales. 1970. On the acquisition of some relational terms. In *Cognition and the Development of Language*, ed. J. R. Hayes, 235–268. New York: Wiley.

Ehri, Linnea C. 1976. Do words really interfere in naming pictures? *Child Development* 47:502–505.

Eilers, Rebecca E., D. Kimbrough Oller, and Judy Ellington. 1974. The acquisition of word-meaning for dimensional adjectives: The long and short of it. *Journal of Child Language* 1:195–204.

Greenberg, Joseph. 1966. *Language Universals*. Janua Linguarum, Series Minor 59. The Hague: Mouton.

Hampton, James A. 2006. Concepts as prototypes. *Psychology of Learning and Motivation* 46:79–113.

Hampton, James A. 2007. Typicality, graded membership, and vagueness. *Cognitive Science* 31:355–384.

Hansen, Nat, and Emmanuel Chemla. 2017. Color adjectives, standards, and thresholds: An experimental investigation. *Linguistics and Philosophy* 40:239–278.

Hare, Richard M. 1952. *The Language of Morals*. Oxford: Clarendon Press.

Jones, Steven. 2002. *Antonymy: A corpus-based perspective*. New York: Routledge.

Jones, Steven. 2007. 'Opposites' in discourse: A comparison of antonym use across four domains. *Journal of Pragmatics* 39:1105–1119.

Jones, Steven, and M. Lynne Murphy. 2005. Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics* 10:401–422.

Justeson, John S., and Slava M. Katz. 1991. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics* 17:1–19.

Kamp, Hans. 1975. Two theories about adjectives. In *Formal Semantics of Natural Language: Papers from a colloquium sponsored by the King's College Research Centre, Cambridge*, ed. E. Keenan, 123–155. Cambridge: Cambridge University Press.

Katz, Jerrold J. 1972. *Semantic Theory*. New York: Harper and Row.

Kennedy, Christopher. 1997. Projecting the adjective: The syntax and semantics of gradability and comparison. Doctoral Dissertation, UC Santa Cruz, Santa Cruz, CA.

Kennedy, Christopher. 1999. *Projecting the Adjective: The syntax and semantics of gradability and comparison*. New York: Routledge.

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

Kennedy, Christopher, and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81:345–381.

Kim, Christina S., and Vilde Reksnes. 2018. Expectations about imprecise language use are speaker-dependent. Presented at the 92^{nd} Annual Meeting of the Linguistic Society of America, Salt Lake City, UT.

Klatzky, Roberta L., Eve V. Clark, and Marlys Macken. 1973. Asymmetries in the acquisition of polar adjectives: Linguistic or conceptual? *Journal of Experimental Child Psychology* 16:32–46.

Klecha, Peter. 2012. Positive and conditional semantics for gradable modals. In *Proceedings of Sinn und Bedeutung 16*, ed. Ana Aguilar Guevara, Anna Chernilovskaya, and Rick Nouwen, volume 2 of *MIT Working Papers in Linguistics*, 363–376. Cambridge, MA: MIT Press.

Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy* 4:1–45.

Konkle, Talia. 2011. The role of real-world size in object representation. Doctoral Dissertation, MIT, Cambridge, MA.

Konkle, Talia, and Alfonso Caramazza. 2013. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience* 33:10235–10242.

Konkle, Talia, and Aude Oliva. 2011. Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human perception and performance* 37:23–37.

Konkle, Talia, and Aude Oliva. 2012. A familiar-size Stroop effect: Real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human perception and performance* 38:561–569.

Kumaran, Dharshan, and James L. McClelland. 2012. Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review* 119:573–616.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82:1–26.

Lasersohn, Peter. 1999. Pragmatic halos. *Language* 75:522–551.

Lassiter, Daniel. 2017. *Graded modality: Qualitative and quantitative perspectives*, volume 10 of *Oxford Studies in Semantics and Pragamtics*. Oxford: Oxford University Press.

Lassiter, Daniel, and Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics & Linguistic Theory (SALT) 23*, ed. T. Snider, 587–610. Ithaca: CLC Publications.

Lassiter, Daniel, and Noah D. Goodman. 2017. Adjectival vagueness in a Bayesian model of interpretation. *Synthese* 194:3801–3836.

Lehrer, Adrienne. 1985. Markedness and antonymy. *Journal of Linguistics* 21:397–429.

Levy, Yonata. 1983a. It's frogs all the way down. *Cognition* 15:75–93.

Levy, Yonata. 1983b. The acquisition of Hebrew plurals: The case of the missing gender category. *Journal of Child Language* 10:107–121.

Lupyan, Gary. 2016. The paradox of the universal triangle: Concepts, language, and prototypes. *The Quarterly Journal of Experimental Psychology* 70:389–412.

Lyons, John. 1977. *Semantics*, volume 2. Cambridge: Cambridge University Press.

Margolis, Eric, and Stephen Laurence. 1999. *Concepts: Core readings*. Cambridge, MA: The MIT Press.

McNally, Louise. 2011. The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In *Vagueness in Communication*, ed. R. Nouwen, R. van Rooij, U. Sauerland, and H.-C. Schmitz, volume 6517 of *Lecture Notes in Computer Science*, 151–168. Berlin: Springer.

Medin, Douglas L., and Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85:207–238.

Mervis, Carolyn B., and Eleanor Rosch. 1981. Categorization of natural objects. *Annual Review of Psychology* 32:89–115.

Muehleisen, Victoria L. 1997. Antonymy and semantic range in English. Doctoral Dissertation, Northwestern University, Evanston, IL.

Murphy, M. Lynne, and Steven Jones. 2008. Antonyms in children's and child-directed speech. *First Language* 28:403–430.

Nosofsky, Robert M. 1987. Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, memory, and cognition* 13:87–108.

Persaud, Kimele, and Pernille Hemmer. 2014. The influence of knowledge and expectations for color on episodic memory. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, ed. P. Bello, M. Guarini, M. McShane, and B. Scassellati, 1162–1167. Austin, TX: Cognitive Science Society.

R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Ravid, Dorit, and Michal Nir. 2000. On the development of the category of adjective in Hebrew. In *From Sound to Sentence: Studies on first language acquisition*, ed. M. Beers, B. van den Bogaerde, G. Bol, J. de Jong, and C. Rooijmans, 113–124. Groningen: Center for Language and Cognition.

Rosch, Eleanor. 1973. On the internal structure of perceptual and semantic categories. In *Cognitive Development and the Acquisition of Language*, ed. T. E. Moore. New York: Academic Press.

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104:192–233.

Rosch, Eleanor. 1977. Human categorization. In *Advances in Cross-Cultural Psychology*, ed. N. Warren, volume 1, 1–49. London: Academic Press.

Rotstein, Carmen, and Yoad Winter. 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics* 12:259–288.

Saryazdi, Raheleh, Julie Bannon, Agatha Rodrigues, Chris Klammer, and Craig G. Chambers. 2018. Picture perfect: A stimulus set of 225 pairs of matched clipart and photographic images normed by Mechanical Turk and laboratory participants. *Behavior Research Methods* 50:2498–2510.

Saryazdi, Raheleh, and Craig G. Chambers. 2018. Mapping language to visual referents: Does the degree of image realism matter? *Acta Psychologica* 182:91–99.

Schmidt, Lauren A. 2009. Meaning and compositionality as statistical induction of categories and constraints. Doctoral Dissertation, MIT, Cambridge, MA.

Schmidt, Lauren A., Noah D. Goodman, David Barner, and Joshua B. Tenenbaum. 2009. How tall is *tall*? Compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, ed. N. Taatgen, H. van Rijn, L. Schomaker, and J. Nerbonne, 3151–3156. Austin, TX: Cognitive Science Society.

Sedivy, Julie C., Michael K. Tanenhaus, Craig G. Chambers, and Gregory N. Carlson. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition* 71:109–147.

Siegel, Muffy. 1979. Measure adjectives in Montague Grammar. In *Linguistics, Philosophy and Montague Grammar*, ed. S. Davis and M. Mithun, 223–262. Austin, TX: University of Texas Press.

Smith, Linda B., Nancy J. Cooney, and Carol McCord. 1986. What is 'high'? The development of reference points for 'high' and 'low'. *Child Development* 57:583–602.

Solt, Stephanie. 2011. Notes on the comparison class. In *Vagueness in Communication*, ed. R. Nouwe, R. van Rooij, U. Sauerland, and H.-C. Schmitz, volume 6517 of *Lecture Notes in Computer Science*, 189–206. Berlin: Springer.

Solt, Stephanie, and Nicole Gotzner. 2012. Who here is *tall*? Comparison classes, standards and scales. In *Pre-Proceedings of the International Conference on Linguistic Evidence*, volume 833, 79–83. Tübingen: Eberhard Karls Universität.

Stolt, Suvi, Leena Haataja, Helena Lapinleimu, and Liisa Lehtonen. 2008. Early lexical development of Finnish children: A longitudinal study. *First Language* 28:259–279.

Syrett, Kristen, Christopher Kennedy, and Jeffrey Lidz. 2010. Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics* 27:1–35.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268:1632–1634.

Tessler, Michael Henry, Michael Lopez-Brau, and Noah D. Goodman. 2017. Warm (for winter): Comparison class understanding in vague language. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, ed. G. Gunzelmann, A. Howes, T. Tenbrink, and E. J. Davelaar, 1181–1186. Austin, TX: Cognitive Science Society.

Toledo, Assaf, and Galit W. Sassoon. 2011. Absolute vs. relative adjectives: Variance within vs. between individuals. In *Semantics & Linguistic Theory (SALT) 21*, ed. N. Ashton, A. Chereches, and D. Lutz, volume 21, 135–154. Ithaca, NY: CLC Publications.

Townsend, David J. 1976. Do children interpret 'marked' comparative adjectives as their opposites? *Journal of Child Language* 3:385–396.

Tribushinina, Elena. 2011. Once again on norms and comparison classes. *Linguistics* 49:525–553.

Tribushinina, Elena. 2013a. Adjective semantics, world knowledge and visual context: Comprehension of size terms by 2- to 7-year-old Dutch-speaking children. *Journal of Psycholinguistic Research* 42:205–225.

Tribushinina, Elena. 2013b. Spatial adjectives in Dutch child language: Towards a usage-based model of adjective acquisition. In *The Construal of Spatial Meaning: Windows into conceptual space*, ed. C. Paradis, J. Hudson, and U. Magnusson, 263–286. New York: Oxford University Press.

Tribushinina, Elena, Huub van den Bergh, Marianne Kilani-Schoch, Ayhan Aksu-Koç, Ineta Dabašinskienė, Gordana Hrzica, Katharina Korecky-Kröll, Sabrina Noccetti, and Wolfgang Dressler. 2013. The role of explicit contrast in adjective acquisition: A cross-linguistic longitudinal study of adjective production in spontaneous child speech and parental input. *First Language* 33:594–616.

Tribushinina, Elena, Huub van den Bergh, Dorit Ravid, Ayhan Aksu-Koç, Marianne Kilani-Schoch, Katharina Korecky-Kröll, Iris Leibovitch-Cohen, Sabine Laaha, Bracha Nir, Wolfgang Dressler, and Steven Gillis. 2014. Development of adjective frequencies across semantic classes: A growth curve analysis of child speech and child-directed speech. *Language, Interaction and Acquisition/Langage, Interaction et Acquisition* 5:185–26.

Tribushinina, Elena, and Steven Gillis. 2012. The acquisition of scalar structures: Production of adjectives and degree markers by Dutch-speaking children and their caregivers. *Linguistics* 50:241–268.

Unger, Peter. 1971. A defense of skepticism. *The Philosophical Review* 80:198–219.

Von Stechow, Arnim. 1984. Comparing semantic theories of comparison. *Journal of Semantics* 3:1–77.

Wheeler, Samuel C. 1972. Attributives and their modifiers. *Noûs* 6:310–334.

Zwicky, Arnold M. 1978. On markedness in morphology. *Die Sprache* 24:129–143.