From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Psychology

2021

When one meaning is not enough: Word learning in a polysemous world

Floyd, Samantha

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01794081302

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

When one meaning is not enough: Word learning in a polysemous world

Samantha Floyd

a dissertation
presented to the faculty
of princeton university
in candidacy for the degree
of Doctor of Philosophy

recommended for acceptance
by the Department of
Psychology

Adviser: Adele E. Goldberg

September 2021

© Copyright by Samantha Floyd, 2021. All rights reserved.

Abstract

Words commonly refer to multiple meanings, a phenomenon known as **polysemy**: *caps* can appear on bottles, human heads, or mushrooms; *green* can refer to a color or an environmental stance; a *sheet* can be in a notebook or on a bed. Polysemy presents a challenge to many theories of meaning and learning, which posit individual meanings and competition between candidate referents. As a result, many open questions remain about how polysemous meanings are learned and represented. In this dissertation*, I address these fundamental issues from a number of angles, using behavioral, developmental, clinical, and computational methods. I first review existing approaches to word meaning and representation (Chapter 2). I then investigate how children and adults uncover and learn the rich, hidden conceptual structure of meanings using novel learning tasks (Chapter 3) and eye-tracking (Chapter 4). In Chapter 5 I investigate variation in generalization by considering a neurodiverse population. In Chapter 6, I develop and evaluate theories of word generalization in quantifiably-testable computational models. Finally, in Chapter 7 I outline a new approach to understanding word learning.

*Portions of this dissertation come from:

(1) Floyd, S. & Goldberg, A.E. (2020). Children make use of relationships across meanings in word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

(2) Floyd, S., Jeppsen, C. & Goldberg, A.E. (2021). Brief Report: Children on the Autism spectrum are challenged by complex word meanings. *Journal of Autism and Developmental Disorders.*

(3) Floyd, S., Goldberg, A.E. & Lew-Williams, C. (2020). Toddlers recognize multiple meanings of polysemous words. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

(4) Floyd, S., Dalawella, K., Lew-Williams, C., Goldberg, A.E. & Griffiths, T. (*Accepted*). Modeling rules and similarity in colexification. *Proceedings of the 43rd Annual Cognitive Science Society Conference*.

Contents

Abstract **iii**

1 Introduction **1**
- 1.1 How are words learned? . . . 3
- 1.2 A new approach to word learning . . . 14
- 1.3 Dissertation overview . . . 15
- 1.4 Chapter summary . . . 16

2 Background **18**
- 2.1 What do words mean? . . . 18
- 2.2 Prototypes, exemplars, and polysemy . . . 20
- 2.3 Chapter summary . . . 27

3 Children make use of relationships across meanings in word learning **29**
- 3.1 Experiment 1 . . . 37
- 3.2 Experiment 2 . . . 47
- 3.3 Chapter summary . . . 62

4 Toddlers recognize multiple meanings—can word learning theories keep up? **64**
- 4.1 Experiment 3 . . . 70
- 4.2 Experiment 3 modeling study . . . 77
- 4.3 Chapter summary . . . 82

5 Children on the autism spectrum are challenged by complex word meanings **84**
- 5.1 Experiment 4 . . . 89
- 5.2 Chapter summary . . . 93

6 Modeling rules and similarity in colexification **95**
6.1 Experiment 5 100
6.2 Experiment 5 modeling study 103
6.3 Chapter summary 110

7 Conclusion: The network perspective **112**
7.1 Chapter summaries 113
7.2 The network perspective 117
7.3 Possible implementations 123
7.4 Summary 128
7.5 To the reader 128

References **145**

Listing of figures

3.1 Polysemy: related senses associated with a single word form (e.g., *cap*) versus ambiguity/homonymy: unrelated senses associated with a single word form (e.g., *bat*). Images publicly available (Barracuda1983, 2006; Crisco, 2014; Delgado, 2012; Siedlecki, 2018; TexasRebel, 2007). 32

3.2 Example set of target objects in the Polysemy condition (one of four sets). The central object represents the prototype. No single feature can successfully discriminate all three targets from foils. 37

3.3 Example set of meanings in the ambiguity condition (one of four sets). Sets were constructed to reduce similarity between meanings. The same "prototype" images were used in both conditions. See the online article for the color version of this figure. 39

3.4 Sample test stimuli in the Label Matching task. Each of the four objects displayed had been labeled by a different word during exposure. See the online article for the color version of this figure. 40

3.5 Average accuracy in the Label Matching task for the ambiguity and polysemy groups. Child performance (4½ -7 years old) at exposure in upper left panel (n = 42 per condition), adults' performance at exposure in upper right panel (n = 42 per condition); 4½ -7-year-old children's performance one week later in bottom left panel (n = 18 in polysemy, n = 19 in ambiguity); adults' performance one week later in bottom right panel (n = 31 in polysemy, n = 28 in ambiguity). Error bars represent bootstrapped 95% Confidence Intervals. . . . 41

3.6 Polysemy trial (one of four). Participants were prompted with a label and asked to select three target meanings. See the online article for the color version of this figure. 50

3.7 Average accuracy in the Sense selection task for the ambiguity and polysemy groups. Child performance (4½ -7 years old) at exposure in upper left panel (n = 42 per condition), adults' performance at exposure in upper right panel (n = 42 per condition); 4½ -7-year-old children's performance one week later in bottom left panel (n = 18 in polysemy, n = 19 in ambiguity); adults' performance one week later in bottom right panel (n = 31 in polysemy, n = 28 in ambiguity). Error bars represent bootstrapped 95% Confidence Intervals. . . . 52

4.1 Schematic of learning in both Cross-situational and Pursuit models. Across three instances, additional meanings become less likely and may become entirely eliminated. 67

4.2 Proportion of looking to target following noun onset (shown at 0ms) in toddlers (*n* = 38) ages 24-36 months on English meaning and Novel extension trials. The window of interest is 300 - 1,800ms following noun onset. 74

4.3 Performance summarized over recognition window for Higher-frequency English meanings, Lower-frequency English meanings, and Novel extensions. . 76

4.4 Each dot represents average performance of one child on each of 6 Novel extension trials. If familiarity with a competing label created competition, driving down performance on Novel extension trials, then the later the word is acquired, the lower performance should be in our toddler participants. However, this pattern was not borne out in the data, suggesting that possible familiarity with the correct English label in Novel extension trials is not responsible for the lower performance. Instead, we suggest that the difference between the English trials and the Novel trials was simply that the children had already become familiar with the conventional meanings while they had not been exposed to the novel meaning. 78

4.5 Model performance on CHILDES input. These results show that learning of single-meaning words, shown in the first panel, is relatively high. The second panel shows the percent of polysemous words for which the models correctly learned at least one meaning. Finally, the third panel shows the percent of polysemous words for which the models successfully learned more than one meaning. 83

5.1 Top row, NT children and children on the spectrum showed equivalent, above-chance accuracy on novel homonyms (pink). While NT children were significantly more accurate on polysemes, children on the spectrum displayed no difference (blue). Bottom row, performance after 1-week delay shows same pattern of results. 94

6.1 A: rule-only model weakly predicts variation in human responses. B: similarity-only model using Sentence-BERT embeddings performs noticeably better. C: a mixture of rules and similarity using Sentence-BERT embeddings correlates even better than similarity alone with the same embeddings. D & E show the same similarity-only and hybrid models, instead using Word2Vec embeddings, which had the same pattern of results (hybrid model > similarity-only). . . 107

Acknowledgments

This dissertation is the product of the support, work, and love of a large community of people.

First, Adele Goldberg: No words can express how you've enriched and contributed to my life. Thank you for all the amazing conversations (yes, even over Zoom!) which reminded me why I love science and kept me inspired all these years. It is difficult for me to say goodbye, because nobody could ever take your place. Thank you for inspiring me, teaching me, welcoming me, supporting me, laughing with me, and always being there. Thank you for providing light when I was lost. Thank you for never trying to create me, but rather helping me to create myself. Thank you for choosing me.

Next, Casey Lew-Williams: Your mentorship and support is like a safe harbor from the world and from my own anxieties. In more than one way, you're the most radical leader I know: you prioritize safety, well-being, and compassion, while offering effective, strategic, and thoughtful feedback. Thank you, from the bottom of my heart. I will never forget how you have made me feel.

Thank you to all my many and incredible mentors: Jill de Villiers, Laura Schulz, Josh Tenenbaum, Lauren Emberson, Tom Griffiths, Jordan Taylor, Uri Hasson, Joan Girgus, Mahesh Srinivasan, Greg Murphy, Ray Jackendoff, Ev Fedorenko, and Ted Gibson. I'll always be grateful for your generosity.

To friends, peers, and collaborators who shared the joy and heartbreak of life and academia with me in roughly reverse-alphabetical order: Zidong Zhao, Xiaofang Yang, Tobias Ungerer, Taylor Martinez, Stephan Meylan, Stats Atwood, Sathvik Nair, Sori Baek, Rachit Dubey, Qihong Lu, Priyam Raghuraman, Nicole Cuneo, Molly Chaney, Mira Nencheva (for the tea and toe rings), Mika Braginsky, Mika Asaba, Mayank Agrawal, Maureen Gill, Maddie Pelz, Madalina Vlasceanu, Lena Skalaban (seriously, thank you), Lauren Feldman, Karan Grewal, Kara Enz, Juliana Trach, Judith Mildner, Joseph Avery, Jordan Starck, Joel Martinez, Jess Mankewitz, Jayden Ziegler, Jason Chin, Jane Smyth, Gabe Moore, Fred Callaway, Eva Fourakis, Emily Liquin, Emily Eyestone, Eliza Kosoy, Crystal Lee, Claudia Yau, Claire Robertson, Chris Crawford, Chelsey Clark, Charlotte Jeppsen, Casey

Lewry, Carolyn Mazzei, Branden Bio, Blake Laham, Benny deMayo, Asana Okocha, Annie Schwartzstein (meet me at the shopping center sign forever please!), Anna Zhukovskaya, Anna Herbolzheimer, Alex Boldin, Ajua Duker, Abla Alaoui-Soce, and Alexia Hernandez.

To those those dear friends from before, and still to this day, thank you from the bottom of my heart: Sonia Brown, Jeremy Hefter, Kiki Broadhurst, David Stoppel, Julianne Boyeon Lee, Eliza Gentzler, David Chou, Matisse DuPont, Katya Raffensperger, and Zarah Sikora.

Thank you to all my more senior friends, peers, and collaborators, for friendship, explaining things to me, and providing support, whether you knew it or not: Zenna Tavares, Yang Wu, Yaelan Jung, Tracy Reuter (special thanks for *so* much help, and for David Attenborough <3), Shahar Shirtz, Jessica Schwab (you're a *star*), Sagi Jaffe-Dax, Robert Hawkins, Paula Rubio-Fernandez, Kate Nealon, Na Yeon Kim, Max Siegel, Martin Zettersten, Lily Tsoi, Libby Barak (for so much! thank you!), Liat Hasenfratz, Julia Leonard, Jessica Kosie, Grace Horton, Gabriel Naiqi Xiao, Felicia Zhang, Elise Piazza, Christine Potter (thank you for answering all my questions and enjoying my most boring proclivities), Bill Thompson, Alexander Paunov, and Amanda Nicol (for putting out the fires).

To all the students at Princeton who taught me so much, thank you: Sonia Ann Friscia, Serena Mon, Sarah Reid, Richard Peng, Richard Yang, Naser Al-Fawakhiri, Michelle Min, Felicia McGill, Matthew Weatherhead, Mariesa Cay, Lily Mannion, Larissa Oliveira (for your brilliant mind and heart), Kennedy Casey, Kavindya Dalawella, Ines Aitsahalia, Ellie Breitfeld, Ella Whitfield, Cierra Moore, Charlotte Brace, Ariella Cohen, Fares Marayati, Annie Jain, and Alice Wang (so excited to call you a colleague!).

To those who gave life to Princeton and made this research possible: RoseMarie Stevenson, Vanessa Gonzalez-Perez, Laura Murray, Shawn Maxam, Jill Ray, Adrian Cupid, Edwin Clayton, Chandra Greenberg, Beth Porter, Keisha Craig, David Carter-Lemos, Jenn Lyons, Ben Singer, Garret McGrath, Randee Tengi, Todd Antonakos, Sarah Lovelace, Jawaan Lowery, Tina McCoy, Danette Quiñones-Silva, Tingting Zhao, Carol Agans, and Gary Drozd.

To my many families, thank you for teaching me a million different ways to love: Jolanta Strojek, Paul Felis, Alex Floyd, Victoria and Jim Klumpp, Genevieve Dionne, Peter and Diana Floyd, Anne Cann, Catherine Bickford, Robert Cann, Rebecca Cann, Anton Lord, Elizabeth Cahn, Fred Cann, Adrian Cann, Elisabeth Piper, Ana Cecilia Jara-Ettinger, Cate Ettinger-McEnulty, Patricio Jara-Ettinger, and Salvador Jara Guerrero.

Wednesday and Fury, you are the best cats. Thank you for loving me.

Thank you to the many people who have been there for me in the ways that matter. Zoe Volenec, the roommate and friend of my dreams, you always help me to be my best self.

Nate Cheek, you always help me to be my most self.

Karina Tachihara, you still have no idea how many times you've saved me from despair

and brought light, joy, and brilliant study designs into my life. My intellectual self is incomplete without you.

Maddie Cann, very few people in the world ever get a best friend anything like you. Your sense of humor, brilliance, selflessness, weirdness, planning and packing ability, responsibility, taste, guidance, cleverness, and compassion is truly one in 7 billion. Thank you for adopting me.

And, last but never least, thank you to the kindest, funniest, smartest, cutest, most fun, fastest typist. Julian, you change my life for the better every moment you're alive. I'm so glad we are quantum entangled, and I love you more than can be expressed (or, "seven").

Consider for example the proceedings that we call "games". I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?... For if you look at them you will not see something that is common to all, but similarities, relationships, and a whole series of them at that... Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ballgames, much that is common is retained, but much is lost.

Ludwig Wittgenstein

1

Introduction

Wittgenstein's* famous example of *game* applies broadly to vocabulary, as the majority of frequently used words† refer to more than a single meaning (Wittgenstein, 1953; Zipf,

*Portions of this Chapter come from Floyd & Goldberg, 2020, Floyd et al., 2020, Floyd et al., 2021b as well as writing submitted in fulfillment of the author's generals examinations.

†Throughout this dissertation, I will refer to "words", "labels" and sometimes "wordforms". The meaning I intend here is general, but is closest to "lemma" or root, and many of the ideas may apply equally to

1945b). Though a smaller proportion (4%) of words across language are **ambiguous** or **homonymous** with multiple, unrelated meanings (such as a baseball *bat* and the mammal *bat*), up to an estimated 80% of words are **polysemous**, in that they refer to multiple meanings which are related to one another (such as baseball *cap*, pen *cap*, and bottle *cap*; Bréal 1897; Dautriche 2015a; Durkin & Manning 1989a; Fellbaum 1998; Navigli & Ponzetto 2012b; Rodd et al. 2002a; Schmitt & McCarthy 1997). Work in philosophy and cognitive linguistics has long puzzled over the multiple meanings of words, documenting systematic patterns (*e.g.* English uses words like *fish* and *chicken* to refer to an animal as well as the animal's meat). At the same time, irregular conventions are also found within and across languages, *e.g.* English uses *beef* instead of extending the label *cow* to the meat of the animal, and other languages prefer to extend their word for cow differently, such as to the meaning "buffalo" (Fillmore & Langėndoen, 1971; François, 2008; Heine & Reh, 1984a; Rzymski et al., 2020). These examples highlight the richness, complexity, and idiosyncrasy of word meanings in natural language. Inspired by these facts, this dissertation will argue that human language learning relies on tracking multiple, distinct meanings and forming generalizations across this network of representations for words.

Cognitive psychologists have investigated a related phenomenon—that naturalistic categories are highly complex. This field has grappled directly with the challenge of concepts which do not fit a single definition or rule (Murphy, 2004a; Rosch, 1978; Rosch & Mervis, 1975; Rosch et al., 1976). As previously mentioned, the philosopher Ludwig Wittgenstein used the example of a category of activities called *games*, which can can be played in groups or alone, with or without defined regulations, competitively or simply as a procedure (*e.g*

morphemes and even to aspects of constructions.

"I'm playing a game where I throw these grapes and try to catch them in my mouth"), or even used to refer to manipulative behaviors in a relationship ("they played games with me and broke my heart") (Murphy, 2004a; Wittgenstein, 1953). Similarly, computational linguists have long faced the challenge of variation in the meanings of words, and have used a variety of techniques to incorporate surrounding context into representations of meaning (Devlin et al., 2019; Griffiths et al., 2007).

1.1 How are words learned?

But despite the ubiquity of polysemy in language, proposals about how children learn words have overwhelmingly ignored the fact that words typically have multiple meanings. Instead, these theories have focused on addressing the challenge of learning a *single* correct meaning for a form. Indeed, since at least 1960, such theories have focused on the learning challenge posed by natural language's *indeterminacy of reference*, often referred to as the "gavagai" problem (Quine 1960, see also Gleitman 1990). In an imaginary scenario, Quine describes a traveler's visit to a foreign land in which the visitor speaks no language in common with the inhabitants, the *Arrernte*, an non-imaginary group of Aboriginal Australian peoples. While standing in a field, a rabbit runs by. One of the inhabitants points and says, "gavagai!". Quine makes the point that the visitor is likely to infer that *gavagai* means "rabbit", but highlights that the environment and language input is underdetermined, and the word may in fact refer to anything:

> "Who knows but what the objects to which this term applies are not rabbits after all, but mere stages, or brief temporal segments, of rabbits? In either event the stimulus situations that prompt assent to 'Gavagai' would be the

> same as for 'Rabbit'. Or perhaps the objects to which 'gavagai' applies are all and sundry undetached parts of rabbits; again the stimulus meaning would register no difference. When from the sameness of stimulus meanings of 'Gavagai' and 'Rabbit' the linguist leaps to the conclusion that a gavagai is a whole enduring rabbit, he is just taking for granted that the native is enough like us to have a brief general term for rabbits and no brief general term for rabbit stages or parts." –Quine 1960

This scenario raises the issue that an interlocutor with no prior knowledge of a language cannot be sure what exactly a speaker is labeling. As a consequence of this influential observation, most developmental work has specifically focused on explaining how children disambiguate between multiple candidate meanings, such as how children determine that *rabbit* refers to a rabbit, rather than the tree nearby. Indeed, this hypothetical challenge to word learning has has motivated the vast majority of existing theories of word learning. The current Chapter will review such proposals and outline several shortcomings, and will conclude by outlining components of a new approach.

This Chapter will start by considering two predominant word learning approaches: Cross-situational word learning and Propose-but-verify (and its recent revision, Pursuit). Because of the ubiquity of words with multiple meanings (Rodd et al., 2002b; Zipf, 1945a), special attention will be paid to the predictions these accounts make for the learning of polysemous words. I will argue that because developmental and computational work on word learning has focused nearly exclusively on one-to-one mappings, it has developed accounts which preclude the learning of polysemy, in which a label refers to multiple, related meanings.

As described in the preceding section, word learning theories have focused on Quine's theoretical challenge: How do people avoid learning the other, incorrect meanings present during the use of a word? To address this, word learning theories have predominantly relied on some type of competitive learning mechanism. Theories have either explicitly assumed or preferred a one-to-one mapping between each word and candidate meaning, with candidate meanings previously encountered being weakened or deleted entirely when not present during the use of a word (Smith et al., 2002; Stevens et al., 2017b; Trueswell et al., 2013b). While such proposals were not intended to preclude multi-meaning learning, they inadvertently legislate multi-meaning words out of learners' grasps.

The strongest version of such a proposal will be described first. "Propose but Verify" (PbV) claims that the word learning process consists of hypothesizing a single meaning, solely from highly informative, unambiguous naming events (Trueswell et al., 2013a). Then, this meaning is either retained in subsequent encounters when it is confirmed (i.e. when the meaning is present during the word's use), or the meaning is struck entirely when disconfirmed (i.e. the word form is heard in a context where available meanings are not consistent with the previously hypothesized meaning). Once a meaning is disconfirmed, an entirely new meaning may be hypothesized in the current or next context, depending on whether the event is highly ambiguous or not.

Though PbV does not provide clear criteria for what determines a highly informative utterance in the real world, the authors demonstrate in separate work the ambiguity within a learning study does influence word learning. Specifically, the relationship between context informativity and reference identification is nonlinear. In a learning experiment where adults were asked to guess the referent of a bleeped naming event, during the course of

a video of a mother-child interaction, only participants who initially witnessed a highly-informative naming event performed at above-chance levels by the end of the study (Medina et al., 2011). The other participants, who were given several trials to guess the referent, did not converge on the correct object at above-chance levels. The authors claim that learners cannot simply combine evidence across exposure, given that study participants who witnessed a highly informative trial following the low-informative trials did not benefit from the preceding ambiguous evidence. This finding is interpreted as evidence that even adult learners may not successfully learn words when ambiguity is high, calling into question the intuition that learners might be able to aggregate possible meanings across experiences to learn words.

It is worth noting that the stimuli used in these studies are created from a separate, third-party perspective to either the parent or child in the scene. This is important as the viewpoint chosen for studies may influence the ambiguity present in a learning scene, and introduce a mismatch between the visual input and a speaker's utterance. In fact, a more recent study took recordings from the child's perspective, using head-mounted cameras, following the intuition that caregivers are more likely to label objects to which a child learner is attending, and that children may have a more constrained visual frame. Consistent with this, participants watching these videos demonstrated significant improvement in accuracy identifying the bleeped word's referent, even after initial incorrect guesses (Yurovsky et al., 2013). These learners were are able to track evidence about a word's referent across multiple instances, even when their initial experience with the meaning was more noisy. This may suggest that learners' difficulties in tracking meanings across multiple contexts in past studies may have been an artefact of the stimuli, which did not reflect the more constrained

input of naturalistic learning.

More relevant to this dissertation, Propose but Verify also predicts that learning multiple word-to-meaning mappings should be impossible. This is because of built-in competition: during exposure, learners can only track a single hypothesized meaning at any point in time. Assuming that a learner first encounters the "bottle cap" meaning for the word *cap*, PbV predicts that this meaning would be struck once the learner hears the label in the presence of a baseball cap. Should the learner next encounter the the cap of a marker, the prior baseball cap meaning would also be deleted. And, because PbV claims that evidence for a different meaning of a word a form is evidence against any preexisting meaning, learners would be equally lost when confronted with ambiguous words which do not share similarity (homonyms or homophones).

A recent variant on this theory has been proposed in the form of a local associative learning model, Pursuit (Stevens et al., 2016). This model is heavily influenced by PbV in that it only represents a single meaning per scene or naming event. However, it deviates from PbV in in two key ways that pertain to the present discussion. First, it keeps a 'count' of the evidence for a hypothesized meaning, in the form of an association based in the the number of times the referent was selected as compared to the possible referents in the learning event. It also specifies a slightly different process upon disconfirmation of a hypothesized meaning: instead of striking the meaning entirely, this model decrements the disconfirmed hypothesis by reducing the strength of the association, and selects a new meaning at random from the scene as the new, dominant hypothesis. This decrement can be set to any value and, in the model implementation, is selected by optimization. Critically, through one or repeated disconfirmed encounters, this decrementation may lead to the same meaning deletion as

PbV (see Chapter 4). Critically, however, Pursuit *in principle* allows for learners to represent multiple meanings for a form.

While Pursuit is, in theory, an improvement for the hypothetical fate of polysemous or ambiguous word learning, the proposal may still fall short in several important ways. It continues to rely on the assumption of competition- that any evidence for a meaning should be evidence against any pre-existing meaning. Relatedly, it does not allow for similarity to reduce possible competition between meanings, *e.g.* that two very similar meanings, such as *chicken* the animal and *chicken* the meat, should compete with one another for their label. Further, this theory would fail to account for the intuition that learners should find polysemous word learning, between related meanings, easier than homonymous word learning, between arbitrary meanings (see Chapter 3).

The assumption of competition between meanings also presents a significant problem for both the PbV and Pursuit approaches in an additional way. Beyond the simple fact that most words have more than a single meaning, most high-frequency words have a large number of senses. Early foundational work by Caramazza & Grober found that a simple word such as *line* has up to 26 distinct, related senses, while the verb *run* has 40 1976:

(1) Ford is coming out with a new *line* of hard tops.

(2) I am no longer in that *line* of business.

(3) He had come from a *line* of wealthy noblemen.

(4) They came to two different conclusions using the same *line* of reasoning.

(5) Sam owned the local bus *line*.

(6) When the curtain rose for the second act, Bob could not recall his opening *line*.

(7) She said it was a *line* from Keats.

(8) He began to type the first *line* of his paper.

(9) When you arrive in New York, please remember to drop me a *line*.

(10) The tailor would *line* the coat with fur.

(11) The rich man was able to *line* his pocket with money.

(12) We wanted to *line* the street with people.

(13) We were told to *line* up.

(14) *Line* your paper for writing.

(15) Sergeant Jones would bring him into *line*.

(16) We built a fence along the property *line*.

(17) There was no turning back; they had crossed the enemy *line*.

(18) The children were playing in the direct *line* of fire.

(19) The judge had to draw a *line* between right and wrong.

(20) I pulled on the *line* with all my strength.

(21) The workman broke through to the gas *line*.

(22) The shortest distance between two points is a straight *line*.

(23) We were told to draw a *line* under the title of the book.

(24) When he frowned, a *line* formed across his forehead.

(25) The road was flanked by a *line* of trees on either side.

(26) As I examined the wall of the cave, I could clearly see a *line* of iron ore.

Were it the case that the 25 meanings had to be decremented each time one in particular was encountered, it is likely that only the single most frequent meaning (perhaps "(13) We were told to *line* up") would be successfully learned, while most other meanings would be at or around zero. However, adult learners are typically familiar with most or all of the meanings listed.

Importantly, it is possible that, rather than compete with one another, a number of of these meanings cluster together based off of similarity, overlap, and learned relationships. But existing models offer no such mechanisms. PbV and Pursuit represent word meanings as atomic, with no features or attributes to provide internal structure. Therefore, these theories' model implementations cannot test the implications of related or overlapping features common between, for example, the sense of *line* in 23 (*We were told to draw a line under the title of the book*) and 14 (*Line your paper for writing*). Of course, this lack of internal structure is unlikely to be a theoretical claim about human concepts, but rather simply a limitation of the proposal. Additionally, while Pursuit may be able to represent additional, decremented meanings, it does not predict that polysemy should be learned more easily than homonymy. If, as the theory states, additional meanings are chosen at random and shared features are not represented or jointly strengthened, children should counterintuitively find learning two related meanings to be as difficult as learning the two entirely unrelated meanings.

Interestingly, several of the same implications are shared by the other predominant word learning approach, Cross-Situational word learning. This theory is often presented in direct

contrast with PbV, as it proposes an associative mechanism which encodes all possible referents for a wordform over many successive instances. Over many experiences with word labels and the surrounding environment, it postits that learners aggregate over all occurrences to strengthen the word's associations to its correct meanings (Hume, 1955; Yu & Smith, 2007). Like PbV and Pursuit, Cross-Situational word learning has amassed experimental evidence from studies with word learners.

In 2008, Smith & Yu showed that infants between the ages of 12 and 14 months could learn novel words from multiple exposures, even when the learning evidence was highly ambiguous within each exposure. In this design, participants heard two novel labels and were shown two novel meanings on each trial, making it impossible for children to determine from a single encounter which was the correct mapping. Indeed, this exposure made it almost impossible for learners to successfully use a single-hypothesis testing mechanism within a given trial. After a block of exposures, children heard one of the labels they had learned, and were shown both the correct referent and another novel meaning from the exposures. Analyses found that babies looked significantly longer to targets than the foil meaning at test, providing strong evidence that, even in babies, learning across ambiguous evidence is possible. Of course, it is still possible that learners used local, within-trial disambiguation, choosing only a single meaning on each trial. However, this account provides a stronger associative foundation for more than one meaning to be tracked across successive learning encounters, offering a possible foundation for learning polysemy.

Cross-Situational learning takes the intersection of both past experiences (other contexts) and the present input to disambiguate reference. Because of this, it is often called a global learning process, as opposed to the local process in PbV or Pursuit, which can only

disambiguate between possible meanings within the immediate context or scene. However, most models of the Cross-situational theory actually rely on competition at both levels (Fazly et al., 2010; Frank et al., 2009; Yu, 2008). In fact, in several experiments, Yurovsky et al. posit that adults use competition at both levels in word learning, treating evidence for a given word-referent pair as evidence against another pairing of the word with a referent 2013. This work describes a mechanism, "in which stronger representations actively inhibit competing representations and thereby make them even weaker," (p. 891). This is argued to ease the learning pressures, since it reduces the size of the lexicon which must be stored. Additionally, it proposes a mechanism by which learners might learn more rapidly than the original, slow association account might predict. In these experiments, adults were shown displays of four objects and heard a stream of words. Some of the words referred to more than one object, which could appear on the screen simultaneously (local competition, Experiment 1), or could appear across multiple scenes (global competition, Experiment 2). Both experiments found that participants could learn multiple referents for words reliably above chance in both global and local learning environments, but that they performed significantly worse on multiple-meaning words than single-meaning words. The authors therefore conclude that, because performance suffered when words were associated with multiple referents, that learners must rely on a one-to-one assumption to use using for one possible meaning as evidence against another. This evidence is useful in understanding that learners may prefer to learn more straightforward words. At the same time, it is limited in its implications for understanding naturalistic word meaning, as the multi-meaning words used in the study were artificial and chosen arbitrarily, while multi-meaning words in naturalistic language strongly favor sharing a label between related meanings.

Further, it is not surprising that adults can employ competition between meanings as a strategy to rapidly succeed in a task constrained by single-meaning words. However, this mechanism still fails to explain how people learn multiple meanings when they are encountered. And while Cross-situational word learning models do allow multiple possible meanings to be represented for a single word, it is unknown whether this in fact mimics how these multiple senses are learned by humans. It seems unlikely that children seriously and independently entertain every possible meaning in a scene each time a word is spoken, as Cross-situational learning assumes. For example, toddlers have been shown to make inferences about their interlocutor's referential intent using information about their physical positioning and eye gaze (Baldwin, 1993).

Relevant to the present questions about multiple meanings, Cross-situational word learning theories are mute on the topic of similarity between meanings, predicting that evidence for any meaning can serve as competition with other candidate meanings. This is largely because Cross-situational word learning, like PbV and Pursuit, treats word meanings as atomic, lacking internal structure which could support the representation of features or attributes for similarity, with each candidate meaning equally dissimilar from its competitors. This presents a potential limitation of the theory. At best, it offers no account for the learning or representation of polysemous words, predicting that learning entirely different meanings for a word should be just as difficult as learning meanings which are related. At worst, mechanisms of competition within the theory predict that lower-frequency meanings should be weakened each time, making an additional meaning potentially impossible to use or recall.

1.2 A new approach to word learning

The prediction that learning multiple meanings should be difficult or impossible, though largely shared in both of the theoretical approaches reviewed above, lacks clear empirical support. Though work has shown that very early learning of homophones can be fragile and dependent on the linguistic context (Dautriche et al., 2018), prior studies have found that children as young as 3 years are familiar with the multiple meanings of English homophonous words, are aware that they share a phonological form, and can successfully categorize their meanings separately (Backscheider & Gelman, 1995). Several studies since have shown that older children are adept at using regularities and patterns to infer additional meanings (Srinivasan et al., 2017; Srinivasan & Snedeker, 2011). However, some word extension regularities, such as the shape bias, have been shown to develop more slowly than word knowledge itself. In particular, the tendency to use shape to guide label extension has been shown to develop in response to noun learning, rather than as a mechanism responsible for early noun learning in the first place (Smith et al., 2002). Therefore the present dissertation addresses two central questions which remain open. First, what kind of word learning process could account for humans' eventual knowledge of multiple meanings? Second, what evidence is there for such a learning mechanism? In response to the first question, I will briefly describe a novel approach to characterizing word learning: The Network Perspective. Then, in the empirical chapters that follow, I will provide evidence for the main pieces of this approach. In the final Chapter I will discuss possible implementations for such an approach.

The Network Perspective proposes that the goal of word learning is to encode networks

of meaning representations which are held together by relationships across each word's meanings. Central to this approach are three main components: distributed representations, additive learning, and generalization. First, distributed representations of meanings provide a basis for meanings to relate to one another. This allows meanings to be represented multi-dimensionally or through multiple features, enabling meanings which are similar, overlapping, or whose features may in part related to cluster together and reinforce one another in semantic space. Second, additive learning allows more than one meaning to be associated with a word. This component of the approach is very simple, but contrasts with existing approaches which overwhelmingly enforce competition between candidate meanings. When used in combination with distributed meaning representations, the additive component of the approach provides an account for how multiple meanings for the same word might reinforce one another on their relevant dimensions, potentially allowing for multiple meanings to be learned rapidly and retained. Finally, the process of generalization across meanings provides an account for how relationships between meanings are formed. The empirical chapters of this dissertation (chapters 3, 4, 5 and 6) will offer evidence for the components, following a review of relevant literature in psychology in Chapter 2.

1.3 Dissertation overview

The present Chapter has reviewed past work on word learning, and outlined several basic components of a new approach. The next Chapter will review past literature on categories in psychology, which will serve as the foundation for this new approach. In Chapter 3, I will present evidence that children and adults learn and retain multiple meanings for novel words when these meanings relate to one another, often in unpredictable, multidi-

mensional ways. In Chapter 4, I present evidence that even toddlers have already learned multiple, polysemous meanings in an eye-tracking experiment, and I use corpora and computational modeling to argue that the predominant theories of word learning fall short in accounting for this knowledge. In Chapter 5, I take a closer look at what constitutes the relationships between polysemous meanings by investigating learning in Autistic youth, a population known to face challenges forming generalizations. In Chapter 6, I examine the role of systematic rules vs. idiosyncratic relatedness in people's expectations about polysemy, using computational models to investigate the joint contributions of both mechanisms to multi-meaning word knowledge in humans. Finally, in Chapter 7, I outline the Network Perspective, my theoretical proposal for learning word meanings, and describe possible approaches to implementing the proposal.

1.4 Chapter summary

Words in natural languages are commonly associated with a group of related, conventional meanings (polysemy). While polysemy is common, little is known about how learners contend with this challenge. Is it possible that they are influenced by relationships among meanings, even when those relationships must be gleaned during the word learning process itself? Existing theories of word learning posit mechanisms of competition between a word's candidate meanings during word learning, and do not provide mechanisms for meanings to relate to one another, suggesting that learners should struggle to learn more than one meaning for a word. This, of course, is at odds with the frequency of words with more than one meaning in natural language. In contrast, I suggest that word learning involves additively encoding networks of distributed meaning representations which are sup-

ported by the relationships learned through generalization.

It is possible that children learn each word meaning independently, in which case learning polysemy would be essentially the same as learning ambiguity. On the other hand, related senses may reinforce each other to some extent during the learning process, predicting that polysemy should be easier to learn than ambiguity (also called homonymy or homophony). Such a fact bears on whether word learning should be studied more as an additive process, rather than a competitive one.

2

Background

2.1 What do words mean?

One approach to word meanings, starting with Aristotle, has studied words as if they could be categories characterized by definitions (Apostle, 1980; Aristotle, BCE)*. This idea is in-

*Portions of this chapter come from Floyd & Goldberg, 2020, Floyd et al., 2020, as well as writing submitted in fulfillment of the author's generals examinations.

tuitive: that there is some attribute or set of attributes in common across all meanings of a word, or all members of a category, which determine the word or category's definition. According to such classical accounts, these attributes are used as a set of necessary and sufficient conditions or formal rules which constitute category membership (Katz & Postal, 1964; Smoke, 1932). This approach predicts that, if a meaning meets this criteria, a label for the category can be used. An avalanche of findings from psychological research complicates this account.

Why can't word meanings be captured by definitions? One reason discovered by Rosch and colleagues is the prevalence of *typicality effects*, or humans' sensitivity to how typical a given item is of its category (Rosch, 1978; Rosch & Mervis, 1975; Rosch et al., 1976). Typicality effects have been shown in reaction time tasks, sorting and rating paradigms, and have robustly replicated across many experiments. In several of these studies, a group of participants rated how typical certain items of an for a label were. A separate group of participants learned novel categories from these examples. These participants were much faster to learn the categories when the examples had been rated as more typical by the first group of participants (Rosch et al., 1976).

Further studies demonstrated that more typical category members were faster to be named by participants (Mervis et al., 1976) and to be identified as category members (Rips et al., 1973). Around the same time, McCloskey & Glucksberg found that many everyday objects, such as olives or curtains, are often categorized and labeled as *fruit* or *furniture* inconsistently, even by the same participants on repeated trials (1978). This finding provided an independent source of evidence against the classical account for words and categories, which predicted that any items fulfilling certain necessary and sufficient conditions should

fall into a discrete category named by the word label. Instead, human categorization is less consistent, far more flexible, and sensitive to the specifics of the concepts being judged.

2.2 Prototypes, exemplars, and polysemy

One way to explain the existence of typicality effects is the existence of an abstract summary representation or *prototype* for determining category membership. The prototype representation has been theorized not to reflect properties of the most frequent category members, but rather reflects the features most in common among category members while limiting the number of properties it has in common with non-members (Rosch & Mervis, 1975). For example, while a "labradoodle" is a relatively low-frequency type of dog, it is a good example of a prototypical dog because it shares many features with other dogs and few features in common with non-dog neighboring categories (e.g., cats or bears). Thus, positing a prototype as a central, informative example can explain typicality effects, as the difficulty of category membership can be thought of as a function of its similarity to the abstract summary representation. Variations on how this summary representation would be implemented have been proposed, as lists or sets of features, feature combinations, or schemas (Hampton, 1979; Medin & Smith, 1981; Murphy, 2004b; Rosch & Mervis, 1975; Rumelhart & Ortony, 1977). Such approaches are able to accommodate different levels of variation across concepts by incorporating features from all category members and weighting them by their centrality to the concept. This has the benefit of taking into account co-occurrences of features and their usefulness in discriminating from the neighboring categories.

Interestingly, however, effects such as typicality which can be interpreted as evidence

for prototype representations, can also be explained by the competing *exemplar theory*. On this account, learners simply store the entire set of encounters with that category, rather than a general, abstracted representation (Hintzman, 1986; Medin & Schaffer, 1978; Murphy, 2004b). To determine a item's category, learners who encounter a novel exemplar can compute the overall similarity of the item to the past experiences stored in memory. This differs considerably from the prototype approach, some versions of which claimed that the concept representation was limited to the summary representations or a weighted list of features. But, like the prototype theory, the exemplar theory also predicts that items which are more typical should be easier to categorize or name, since typical items are more likely to share greater similarity with other members of the category and less similarity to members of competing categories (Mack et al., 2013; Medin & Schaffer, 1978; Smith & Minda, 1998; Zaki et al., 2003).

It is important to note that the nature of typicality effects themselves has been debated. While is clear that the effects of typicality are indeed real and robust, the representations they are argued to implicate remain unclear. Specifically, psychologists have argued that typicality effects may be epiphenomenal, as even concepts which are formally defined, such as *odd number* or *triangle*, can show graded judgments from humans according to typicality. Given that both concepts which are definable (such as *prime number* or *triangle*) and concepts used to argue for summary representations (such as *bird* or *furniture*) may show graded typicality effects to the same degree, this work suggests that such effects are not evidence for a different underlying representation (Armstrong et al., 1983). Alternatively, it is possible that particular exemplars may cluster together, offering the opportunity to easily calculate and determine typicality across items based on a dense cluster. This, much

like exemplar approaches, would allow for typicality-like judgments, without a separately represented summary representation.

However, subsequent studies have persisted in adjudicating the debate between prototype and exemplar theory, with most laboratory-based studies finding evidence for exemplar effects (McKinley & Nosofsky, 1995; Nosofsky et al., 1994). Is there still a place for prototype representations? Real-world concepts are learned in naturalistic settings, which are complex and may be more challenging to store and represent exemplar-style. Indeed, the more straightforward representation that prototypes offer has been suggested to fare better in learning categories from naturalistic input (Battleday et al., 2019), suggesting that the simplified representations offered by prototype theory may by relied on in everyday cognition.

Given the abundance of work both on prototype- and exemplar-based categorization, it is natural to ask whether there is a category learning theory in which both processes are at play. One possible proposal which may explain how and why stored experiences (such as exemplars) and generalizations (such as in prototypes) co-exist in word and category knowledge is *reminding-based generalization* (Ross et al., 1990). This theory assumes that exemplars are stored, and that learners begin categorizing purely by similarity to past exemplars. However, as learners are repeatedly reminded of stored exemplars, they begin to compare related features to one another. Ross et al. argue that these repeated comparisons are represented and stored as generalizations, and can subsequently be used to categorize novel exemplars in addition to the original, exemplar-based reminding process. Evidence for these processes comes from four experiments. In the first experiment, the authors provided participants two exemplars for each category followed by one of two possible third exemplars.

CATEGORY X:

Exemplar A: Shirley. (1) likes ice cream; (2) goes to the store to buy nails

Exemplar B: Julia. (1) likes to read fiction; (2) goes to the store to buy swimsuit

Test Exemplar: (1) likes sherbet OR fantasy novels (counterbalanced)

(2) buys wood & a towel

In this item, the novel test exemplar shares a feature with one of each of the exemplars of Category X, and partially shares feature two with each of the exemplars from the category. The reminding-based learning theory predicts that, if a participant got the sherbet feature information and categorized the test item into X based on its similarity to Shirley, they would also be more likely to notice the similarity between "buys wood" and "buys nails" (even though they could potentially observe an equally valid similarity between the test exemplar and Julia, based on "buys a towel" and "buys a swimsuit"). In other words, when the test exemplar reminds participants of a specific prior exemplar (Shirley), they compare "wood" and "nails", rather than the equally-possible comparison between "towel" and "swimsuit". The theory then predicts that the generalization between wood and nails is represented and stored, and available for later use. At test, participants rank ordered a list of items based on how likely members of Category X were to purchase them, including two items related to those purchased by category members (sunglasses and a chisel). As predicted, participants rank significantly higher the item which is similar to the generalization drawn between the novel item and the reminder exemplar (in this case, the chisel) instead of the alternative generalization for which they had equal evidence (the similarity between the swimsuit and towel).

The authors argue that neither exemplar nor prototype theories can account for this ef-

fect on their own: early exemplar accounts would simply predict that the commonalities between ice cream + sherbet, nails + wood, and towel + swimsuit would each be calculated equally and should not have greater relative importance to one another, while early prototype theories would predict that both the abstractions about woodworking and beach-going would be stored equally within the summary representation, without the information about which other generalization co-occurred within the same exemplar.

It is possible, of course, that some variation of the exemplar theory could account for this data. One possibility is that, as stored exemplar features are activated through their similarity to the test item, the exemplar as a whole is reactivated, resulting in spreading activation to the co-occurring features of that exemplar. The authors provide evidence that their effect is not solely attributable to this explanation in the next experiment. In this version, participants learned only one feature of the new exemplar which was similar to exemplar for the category. Instead of being similar to features of each of the stored exemplars, the test item's secondary features were changed to be irrelevant (see the example below).

CATEGORY X:

Exemplar A: Shirley. (1) likes ice cream; (2) goes to the store to buy nails

Exemplar B: Julia. (1) likes to read fiction; (2) goes to the store to buy swimsuit

Test Exemplar: (1) likes sherbet OR fantasy novels (counterbalanced)

(2) buys a parakeet and peas

In this example, only one feature of a stored exemplar (Exemplar A, ice cream) is related to a feature of the test exemplar (sherbet). As in the first experiment, after categorizing the test exemplar in X, participants rank the likelihood that a member of that category would buy items which relate to the secondary features of each exemplar (buys wood, buys a

towel). Reminding-based learning predicts that, without a feature in the test exemplar to remind participants of the "nails" feature in Exemplar A, no generalization about "nails" can be formed. Thus, when a nail-related item is ranked for its likelihood to be purchased by someone in the category, it's ranked equally with a feature of the other exemplar of the category (in this case, swimsuit). However, if activating a feature of an exemplar is enough to reactivate all the features which co-occur with it, then the effect from the first experiment should still hold in this design. As predicted and in accordance with the reminder-based learning account, the result was null; without a comparison between nail-related features of the old and new exemplar, no generalization about nails was used to rank nail-related features as more important for the category.

Though these stimuli are organized within fairly artificial categories, such findings hold promise for understanding naturalistic word relationships in polysemy, and challenge traditional notions of categorization. Ross et al. demonstrate that relationship structures could be learned for distinct items which relate to one another. On this theory, exemplars first organize around pure similarity. However, as exemplars repeatedly co-activate, learners engage in a process of comparison between these co-activations. The comparison is then encoded as an abstraction, which can then be generalized to the new input (test exemplar). The theory allows for a gradient of association, where one stored exemplar is the most similar to the test exemplar, though the additional stored exemplars can remain within the category.

Slightly complicating the picture, studies have claimed that polysemous word meanings are not easily explainable through intuitive understandings of categorization (Klein & Murphy, 2001,0). For instance, while an animal's *mouth* bears some resemblance to the

mouth of a river or the *mouth* of a cave, it might be a challenge to argue that these are the same concept or category in an intuitive sense. Experiments offer empirical confirmation for this concern: participants did not categorize meanings of the same polysemous word together in an explicit sorting task (Klein & Murphy, 2002). To ensure this was not simply an artefact of explicit sorting, which could invite participants to differentiate items, the same authors also used a standard implicit categorization task. Based on prior studies, called category-based induction tasks, participants are taught novel properties about individual items which they had limited prior knowledge of, such as types of rocks or animals. Participants are then asked whether the property was also true of other, similar items. Findings confirmed that participants generalize the novel property to items they considered to be the same category, and that participants generalize the property more widely within the category when the novel property is true of a more typical item as compared to an atypical item (Rips et al., 1973). To investigate whether people categorize polysemous meanings together (such as wrapping *paper* and *the paper* as an organization), Klein and Murphy tested how properties were generalized across these meanings. For example, participants were shown the following stimuli:

Suppose that scientists find the *biotin* bacteria in:

wrapping PAPER

Type in the probability (out of 100) that the *biotin* bacteria will also be in:

liberal PAPER

Klein and Murphy compared the strength of a property's generalization between polysemous word pairs to a baseline measure of that property's likelihood to be generalized across

homophones (such as *baseball bat* and *flying bat*). They used a range of property types, as past research has demonstrated that the nature of the property being generalized can have a strong effect on the induction (for instance, biological properties are likely to be generalized more widely for animate creatures than artefacts regardless of the item's category status). They found that participants were not significantly more likely to generalize properties across polysemous word meaning pairs than the baseline of unrelated word pairs. Cognitive linguists, too, have found that polysemous extensions are often idiosyncratic and can span widely across our intuitive understanding of a category (Lakoff, 1987).

As described described above, when categorization is operationalized as a more simplistic, single-dimension extension (*paper* is writing paper when it's made of the same material; chairs are of the same category so long as they serve the same function), it is difficult to use such a framework to make sense of naturalistic polysemy. But, as reviewed earlier, categories may also function at a level above these straightforward extensions. For example, in Ross et al., Shirley, Julia, and the test exemplar all related to one another with slightly different levels of similarity. It is possible that, even though polysemous meanings are not solely predictable based on a prior understanding of categorization, that a richer categorization structure can be applied across meanings to account for both the distinct items (meaning) and their relationships. I will return to such a proposal in Chapter 7.

2.3 Chapter summary

Consistent with Wittgenstein's early observations, empirical work points to the difficulty of using straightforward understandings of categories to account for word meaning. Most relevant to the present dissertation, past work demonstrates that a single-dimension cat-

egory does not suffice in capturing the full range of rich multi–concept meanings in polysemy. The following chapters will explore this richness of word knowledge during the process of learning.

3

Children make use of relationships across meanings in word learning

Work exploring children's knowledge of multiple meanings has been limited, and as a consequence it is not clear at what age children start learning multiple meanings.

This unknown fact bears on critical conclusions about the nature of learning and language: Are additional meanings part of the representational foundations of meaning? Or are additional meanings simply added later, outside of a central or core meaning? Initial evidence has shown that older children are able to take advantage of productive or"regular" polysemy (Rabagliati et al., 2010; Srinivasan et al., 2017,0; Srinivasan & Snedeker, 2011). That is, when a novel word is illustrated with one sense, children may expect the word to apply in a way that is systematically related. For instance, a word that refers to a certain material can be extended to apply to an object made of that material (as exists for the English words, *glass* and *tin*; Srinivasan et al. 2019); a word that refers to an animal can be extended to apply to the meat of that animal (as exists for *chicken*; Srinivasan & Snedeker 2014; and the word for a container can be used to refer to the container's contents (e.g., *wash* vs. *pour* a bowl; Rabagliati et al. 2010). *

In these existing studies, the type of regular pattern between the familiar and secondary meaning was taken from English, as exemplified by the instances provided. Therefore, the polysemous relationships tested were likely already exposed to children. This leaves unaddressed the question of how children contend with multiple meanings before relationships are known to them. A further limitation of the existing work is that many words' meanings in natural language are often not reliably predictable on the basis of productive generalization or regular patterns (Fillmore & Langendoen, 1971; Rabagliati & Snedeker, 2013). For example, while *pen caps*, *cap and trade programs*, and *mushroom caps* are all called *caps*, the relationships among these meanings are not straightforwardly definable and do not recur in the same way for other words (*e.g.* we do not just as frequently talk about *lid and trade*

*Portions of this chapter come from Floyd & Goldberg, 2020 and writing submitted in fulfillment of the author's generals examinations.

programs or *mushroom lids*). Nor is the apparent similarity among the meanings of *cap* itself sufficient to predict that the meanings will share a label; a *baseball cap* is called *gorra* in Spanish, while a *pen cap* is called *tapa* (*de lapicero*). Cross-linguistic differences such as these suggest that speakers often must learn the range of meanings individual words conventionally allow in their language (Bracken et al., 2017; Lehrer, 1990a; Malt, 2010; Murphy, 2004a; Pinker, 2007; Pye, 2017). Indeed, even the patterns that we might consider regular, such as extending a label based on a shared shape across meanings, are known to develop through exposure: as toddlers learn English nouns which tend to generalize based on shape, they subsequently *develop* a propensity to generalize novel noun labels based on the same feature (Smith et al., 2002). This suggests that relationships across meanings of a form are not a clear explanation of how children could learn polysemy, as they themselves must also be learned.

The learners' task would certainly be easier if the meaning of each word in any language could be captured by a single definition or set of necessary and sufficient features. However, as Wittgenstein first pointed out, the range of conventional meanings that common words allow is typically not captured by a single overarching definition (1953). For example, we might attempt to define the English word *cap* in a general way as "something which tightly covers the top of something". But this potential definition does not accurately predict when the word *cap* is appropriate and when it is not; it would incorrectly predict corks, lids, and roofs could be called *caps*, and it would not predict *mushroom caps* or *cap and trade*, as they do not tightly cover anything. In fact, words are routinely extended along multiple dimensions which only loosely cluster together according to family resemblance structures (Rosch & Mervis, 1975; Wittgenstein, 1953), radial categories in which a central

Figure 3.1: Polysemy: related senses associated with a single word form (e.g., *cap*) versus ambiguity/homonymy: unrelated senses associated with a single word form (e.g., *bat*). Images publicly available (Barracuda1983, 2006; Crisco, 2014; Delgado, 2012; Siedlecki, 2018; TexasRebel, 2007).



meaning is extended along more than one different dimension (Lakoff, 1987), or "chaining" from one sense to another (Heine, 1992; Ramiro et al., 2018a). The fact that a word's range of conventional senses is not predictable from a definition or single summary suggests that children must learn, rather than simply predict, additional senses of many words from witnessing the conventions of their language. But, to date, it has not been established that children easily learn the kinds of multidimensional and idiosyncratic relationships that commonly underlie the structure of conventionally polysemous meanings in language (*e.g.*, *cap*).

In the following experiments, we introduce participants to novel words and novel (unfamiliar) meanings in order to control for prior knowledge, including possible metaphorical

or functional interpretations. The meanings are represented by images of novel objects in order to provide well-controlled and brief exposure, as is common in word learning experiments (Golinkoff et al., 1992; Horst & Hout, 2016; Landau et al., 1998). By using novel words and novel meanings, we avoid potential transfer effects from the language children already know, and systematically avoid allowing all meanings of each word to share the type of productive extension that are common cross-linguistically and which have been investigated by others (Srinivasan et al., 2017; Srinivasan & Rabagliati, 2015; Srinivasan & Snedeker, 2011). This is critical for truly determining whether learners track multiple novel word meanings for individual words from the introduction of a new word, and whether learners can spontaneously make use of relationships between the meanings.

There is reason to hypothesize that novel words with multiple meanings will be learned more easily if the meanings are related to one another, stemming from two independent lines of work: work on lexical access in adults and work on non-linguistic categorization. Some of the earliest work on lexical access investigated whether all senses of ambiguous words were simultaneously activated (Swinney, 1979; Tabossi et al., 1987). To the extent that polysemy was recognized, it was suggested that listeners selectively activated an "underspecified" or vague meaning comprised of only features shared by all meanings, allowing listeners to avoid committing to a more specific interpretation until after disambiguation from context (Frazier & Rayner, 1990). In fact, words are recognized faster in lexical-decision tasks when they have multiple related meanings, whereas multiple unrelated meanings do not confer the same advantage (Armstrong & Plaut, 2008; Azuma & Van Orden, 1997; Brocher et al., 2017; Klepousniotou & Baum, 2007; Rodd et al., 2012,0). Rodd (2004) has modeled the distinction between ambiguity and polysemy by positing a dis-

tributed representation of a word's meanings in which unrelated meanings interfere with one another during retrieval in a way that related meanings do not. If, in a parallel way, only ambiguous meanings interfere with one another in the initial stages of learning, we predict that it should be easier to learn and remember polysemous words when compared with ambiguous words.

Certain classic work on non-linguistic categorization also suggests that related meanings of a word should be easier to learn than unrelated meanings of a word. Rosch & Mervis (1975) found that exemplars that shared many features with other members of a category and few features with members of other categories (more prototypical exemplars) were easier to learn than less prototypical exemplars of the same category. At the same time, this work was performed on categories with structure that participants were already familiar with (e.g., furniture, birds, colors). Of current interest is the extent to which a new, latent complex structure will be useful to children during the initial learning of novel words. To ensure that perceptual similarity alone does not determine the requisite relationships required to learn the novel words in our paradigm, we report a Norming study below.

An additional issue raised in recent studies is that comprehension of newly learned novel words appears to degrade very quickly, particularly when participants are exposed to multiple candidate meanings at once (Aravind et al., 2018; Horst & Samuelson, 2008). This raises the possibility that participants simply rely on a strategy to satisfy experimental demands without retaining the meanings of new words (Trueswell et al., 2013b), and is worrisome insofar as real-world vocabulary learning requires maintenance in long term memory. Therefore, to test whether multi-meaning words are retained by participants, we re-test all available learners after a delay of roughly a week.

The present studies are the first that we know of to empirically compare the learning of polysemous and ambiguous words in a way that satisfies several desiderata. We present participants with stimuli that require them to learn word meanings that involve extensions across multiple dimensions, ensuring that no one single feature can be used to identify the meanings (this is confirmed by the additional Norming task). Both novel labels and novel senses are used in order to mitigate the effects of previous experience and world knowledge, and no feedback is provided. To increase the ecological validity of the results, multiple novel words are learned, and we retest participants after a week delay and without re-exposure to investigate retention in long-term memory. We test 4 ½ -7 year old children as well as adults, since children represent a more naïve group of participants who are less likely than adults to rely on metalinguistic strategies (Gombert, 1992) or metacognitive skills (Flavell et al., 1981), reducing the likelihood that their responses reflect learned strategies. Additionally, a number of findings in category and feature learning show young children are substantially more likely to attend widely to features, rather than categorize based on a single dimension as older children and adults tend to (Deng & Sloutsky, 2016; Plebanek & Sloutsky, 2017; Smith & Kemler, 1977), indicating that young children may be well suited to learning the kinds of multidimensional word meanings that polysemous word learning requires. Finally, during this age range, children are known to learn 5-7 words per day (Cunningham, 2005), so the possibility of a polysemy over ambiguity advantage would have implications for our understanding of lexical networks as they are being formed.

We recognize that in reality, relatedness between meanings falls on a cline; while some senses are highly related, others share fewer features, and appear closer to ambiguity (Geeraerts & Gentner, 1993; Tuggy, 1993). Therefore, in the two experiments described below,

meanings for each word in a **Polysemy** condition were constructed to share similarities along two dimensions, while meanings in an Ambiguity condition were constructed by scrambling polysemy sets to reduce similarity between meanings. In a between-subjects design used in both experiments, adults and children were randomly assigned to either the polysemy condition or the ambiguity condition. During exposure, participants witnessed exposure videos which displayed a series of novel labels and pictures of novel objects ("meanings"), labeled with novel words, with pictures of foil objects interwoven.

In order to test whether participants successfully mapped the word forms to their intended targets, in a **Label Matching** task, Experiment 1 presented participants with four targets, each corresponding to a different word; participants were then asked to select the appropriate target for a particular word. To determine whether participants in fact construed multiple polysemous targets as targets of a single label and whether they were able to discriminate those target meanings from foils, a Sense Selection task (Experiment 2) asked participants to identify, for each of the 4 novel words, the 3 targets for each word, distinguishing them from 5 foils, which were witnessed but unlabeled during exposure. Thus, we determined whether participants were able to discriminate one target meaning from distractors which had been labeled differently (Exp. 1), and whether they could group commonly-labeled targets together in the presence of also-witnessed foils (Exp.2). All available participants were retested on the same two tasks after a week-long delay with no intervening exposure.

3.1 Experiment 1

As our interest is in vocabulary learning, Experiment 1 ("Label matching") aims to determine whether adults and children were able to successfully match a novel label with its target meaning by selecting the target meaning from among 3 other meanings which had been labelled differently.

Figure 3.2: Example set of target objects in the Polysemy condition (one of four sets). The central object represents the prototype. No single feature can successfully discriminate all three targets from foils.



Method

Participants. 84 children between ages 4½ - 7 were recruited to the lab or at local elementary schools (M = 5:11; SD = 0.62) and 84 adults were recruited from Amazon's Mechanical Turk through Turk Prime (Litman et al., 2017b). An additional six children's data was excluded from analysis for technical malfunction (N = 2), experimental error (N = 2) or missing date of birth (N = 2), and 3 adult participants' data was missing due to technical malfunction. Sample size and analyses were preregistered for both adults and children at the first timepoint. Participants were randomly and equally assigned to the polysemy or ambiguity condition. Children's ages were matched across conditions (Polysemy: M = 5.88, SD = 0.67; Ambiguity: M = 5.90, SD = 0.57; ß = -0.02, t = -0.375, p = 0.708). Longitudinal analyses were not preregistered because we could not anticipate the rate of attrition, and thus sample size, prior to data collection. Thirty-seven children were available to be

tested again after the 1 week delay (mean age = 6:2, N = 18 in polysemy condition, N = 19 in ambiguity). Fifty-nine adults were also re-recruited after the delay (N= 31 in the polysemy condition, N= 28 in the ambiguity condition). All procedures were approved by the Princeton University institutional review board (approval number 49511). Before starting the study, adult participants and parents/guardians of child participants gave written consent.

Stimuli. In the polysemy condition, one "prototypical" meaning shared a distinctive feature with a second target and a different distinctive feature with the third target, reducing the possibility that a single feature could distinguish all three target objects from foil objects. In Figure 3.2 with targets from the polysemy condition, the object in the center is the prototype; it shares a handle with the object on the left and a material with the object on the right, while the objects on the left and right share no distinguishing features with each other. For each of the 4 novel words, 5 foil "meanings" were also created. Thus, participants in the polysemy condition witnessed 4 novel words, each assigned three novel, related meanings (for a total of 12 target meanings); the videos also contained five novel foil meanings for each word (a total of 20 foils). Stimuli were normed to determine the extent to which visual similarity alone could predict results by exposing a separate group of participants to the same exposure videos without sound (see Norming in Experiment 2 for details).

Participants in the Ambiguity condition heard the same 4 novel labels and witnessed the same 12 target meanings and the same 20 foils, but in this condition the 12 target meanings were scrambled across the sets constructed in polysemy so that the 3 meanings of each word were relatively unrelated to one another (See Figure 3.3).

Figure 3.3: Example set of meanings in the ambiguity condition (one of four sets). Sets were constructed to reduce similarity between meanings. The same "prototype" images were used in both conditions. See the online article for the color version of this figure.



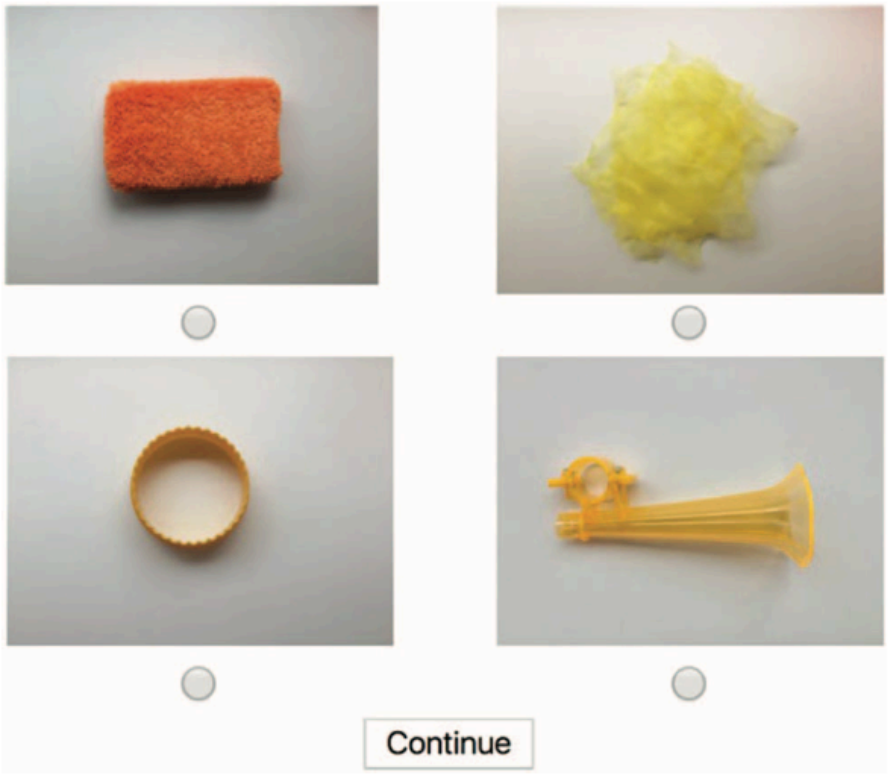
Exposure. Each of two exposure videos witnessed by each participant contained a stream of novel objects, presented one at a time for 2.5 seconds each. All target objects were named once by a human voice ("This is a kaisee", labels counterbalanced), except the prototypical target object, which was shown and named twice. Ten foil images in each video were accompanied by the sound of a bell tone. The ambiguity sets were not constructed to share similarity, making it impossible to select a prototypical meaning, so one meaning of each ambiguous set was chosen at random to appear twice in the video to match exposure across conditions. Each of the 2 videos exposed participants to two novel words, each one of which labeled 3 novel target objects, offering participants in each condition the opportunity to learn 4 novel words with 12 meanings. Each video was approximately 55 seconds long, with targets and foils presented in a fixed, pseudo-randomized order, and the two videos in each condition were presented once each, in counterbalanced order.

Procedure. Web scripts prevented participants from re-watching any exposure videos, and participants could not rewind to extend exposure or skip through. Children were tested in a quiet area, using noise-canceling headphones. The experiment was formatted identically for children and adults, but was shown to children on an external monitor while the experimenter sat to the child's left, also facing the screen. Children selected answers us-

ing a wireless mouse, except if unable to operate it, in which case the experimenter selected the options which the child indicated by pointing.

Test On each of six trials, participants were prompted with one of the words they had been exposed to, and were shown four meanings, one from each of the 4 different novel words they had witnessed. See Figure 3.4. The meanings that had been witnessed twice in the videos (the "prototypes") appeared on the same trials, to control for the amount of exposure. Each participant was tested on all senses of two of the four words they had witnessed. Word assignment was randomized across participants.

Figure 3.4: Sample test stimuli in the Label Matching task. Each of the four objects displayed had been labeled by a different word during exposure. See the online article for the color version of this figure.

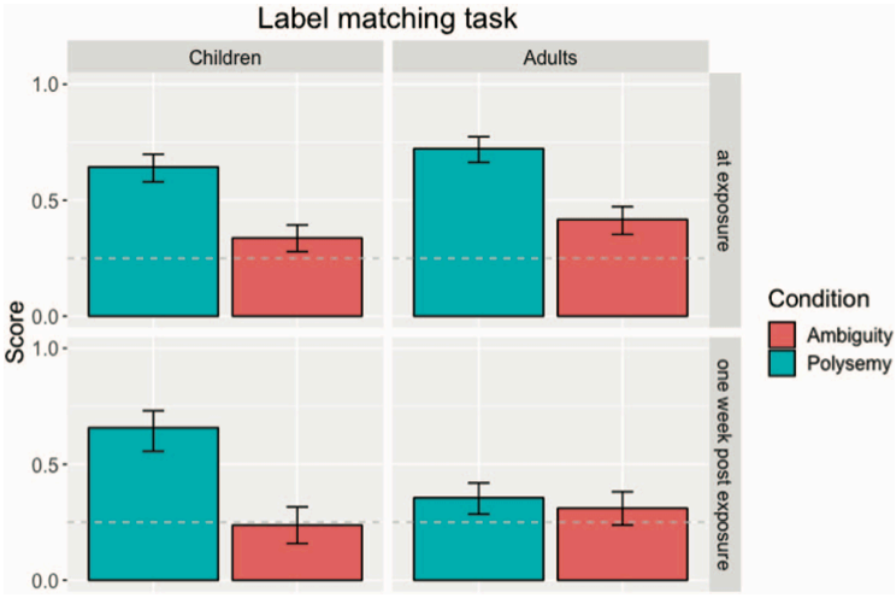


We entered adult and child results from each trial into separate mixed-effect models with

condition (polysemy vs. ambiguity) as the fixed effect using RStudio and the lmerTest library and fit the maximal random terms that convergence would allow (Barr et al., 2013; Kuznetsova et al., 2017; RStudio, 2017). For all models, we preregistered the maximal random effect structure that convergence would allow.

Adults. Figure 3.5 shows performance in the two conditions for adults immediately after exposure in the top right panel. We analyzed results with a multilevel model which included condition as the fixed effect with random intercepts for subjects and items and a random slope and intercept for order. This model revealed significantly better accuracy in the polysemy condition in comparison to the ambiguity condition (ß = 0.31, t = 4.14, p = 0.002).

Figure 3.5: Average accuracy in the Label Matching task for the ambiguity and polysemy groups. Child performance (4½ -7 years old) at exposure in upper left panel (n = 42 per condition), adults' performance at exposure in upper right panel (n = 42 per condition); 4½ -7-year-old children's performance one week later in bottom left panel (n = 18 in polysemy, n = 19 in ambiguity); adults' performance one week later in bottom right panel (n = 31 in polysemy, n = 28 in ambiguity). Error bars represent bootstrapped 95% Confidence Intervals.



Children. Figure 3.5 shows children's performance at the time of exposure in the upper left panel. As with adults, the preregistered multilevel model included condition as the fixed effect and random slopes and intercepts for order and subjects, and random intercepts for item. The model revealed significantly better accuracy in the Polysemy condition compared to the Ambiguity condition (ß = 0.31, t = 5.10, p = 0.002).

Longitudinal results. The longitudinal analyses were not preregistered because we could not anticipate available sample sizes. The bottom panel in Figure 3.5 shows participants' accuracy averaged in each condition at the second timepoint. The children who returned for the retest did not differ in age across conditions, (polysemy: M = 6.09, SD = 0.38; ambiguity: M = 6.17, SD = 0.36; ß = -0.07, t = -0.61, p = 0.54). Re-tested children returned after an average of 7.19 days (range = 7-8 days, SD = 0.4 days), and this length was not significantly different across the conditions (polysemy: M = 7.17 days, SD = 0.38; ambiguity: M = 7.21 days, SD = 0.42; ß = -0.04, t = -0.33, p = 0.74). Adults were re-recruited using Turk Prime (Litman et al., 2017b).

Children. To first determine whether the returning child participants were representative of the full child sample, we entered first timepoint results into a multilevel model to compare performance at exposure between the whole group with the returning subgroup, using Group, condition, and their interaction as fixed effects, and random intercepts and slopes for order, item, and subjects. This revealed no main effect of Group (full group) (ß = -0.03, t = -0. 45, p = 0.66) and no interaction of condition (polysemy) and group (ß = -0. 02315, t = -0.30, p = 0.76), but again revealed a main effect of the polysemy condition (ß = 0. 30, t = 4.2, p = 0.001).

The maximal converging model at the second timepoint included random intercepts and

slopes for participants, items, and order, as well as condition as the fixed effect. Children again performed significantly better in the polysemy condition at the second time point (ß = 0.42, t = 6.63, p = 2.08e-08). To determine if there were changes in accuracy across timepoints, we analyzed just those children who participated at both timepoints, using a model with timepoint, condition, and their interaction as the fixed effects and subjects, order, and items with random intercepts and slopes. This revealed a main effect of condition (ß = 0.33, t = 3.987, p = 0.007), and a main effect of timepoint (T2) (ß = -0.15, t = -2.38, p = 0.02) but no significant interaction of condition (polysemy) and timepoint (T2) (ß = 0.09, t = 0.77, p = 0.47).

Adults. To determine whether the returning adult participants were representative of the full adult sample, we entered group (full vs. returning subgroup), condition, and their interaction as main effects (and random intercepts and slopes for order, item, and subjects). This revealed no main effect of group (returning subgroup) (ß = 0.01, t = 0.24, p = 0.81) and no interaction of condition (polysemy) and group (ß = 0.00, t = 0.00, p = 0.99), but again revealed a main effect of the polysemy condition (ß = 0.29, t = 3.379, p = 0.003). We also fit the maximal converging multilevel model to predict adult performance using condition as fixed effect, with random intercepts and slopes for subjects, items, and order. Unlike both the adult data at the first timepoint, and the child data at the first and second timepoints, adults' polysemy advantage at the second timepoint did not approach significance (ß = 0.05, t = 0.6, p = 0.56).

Finally, we analyzed across timepoints, including just those participants who participated at both timepoints (n = 59), again using a maximal converging model with timepoint, condition, and their interaction as fixed effects (random intercepts and slopes for subjects,

item, and order). This analysis revealed a significant main effect of the polysemy condition (ß = 0.31, t = 4.21, p = 0.0002), but only a marginal effect for timepoint (ß = -0.14, t = -1.91, p = 0.06), and a significant interaction between time and polysemy (ß = - 0.26, t = -2.76, p = 0.007). The interaction may suggest that though performance was lower overall for adults at the second timepoint, adults forgot significantly more in the polysemy condition. Alternatively, it may be due to the floor effect for ambiguity at the second timepoint: i.e. for performance in the ambiguity condition to decay at the same rate as the polysemy condition, it would have had to be reliably below chance at timepoint two (see Comparisons to chance).

Age comparisons. Our preregistered analyses reveal that the polysemy over ambiguity advantage was present both in children and in adults at the first timepoint, and an exploratory comparison between age groups allowed us to investigate whether it was significantly different at either age. We first compared performance at time of exposure, fitting a model with age group, condition, and their interaction as fixed effects and once again, the maximal converging structure including a random intercept for subjects and random slopes and intercepts for items and order. The model revealed a significant effect of condition (polysemy) (ß = 0.30, t = 5.62, p = 2.12e-07), but no significant effect for age group (children) (ß = -0.08, t = -1.52, p = 0.13), and no significant interaction (ß = 0.004, t = 0.05, p = 0.96), suggesting that both adults and children exhibit the polysemy advantage to roughly the same degree at the first timepoint. We also compared group performance across conditions at the second timepoint, and found a significant interaction between condition and age group (ß = 0.38, t = 3.78, p = 0.001), but no main effect for condition (ß = 0.05, t = 0.56, p = 0.59) or age group (ß =-0.07, t = -1.24, p = 0.23). This suggests that, unlike

timepoint 1, in which both adults and children showed a significant polysemy advantage, children at the second timepoint performed significantly better in the polysemy condition than adults, who showed little evidence of a polysemy advantage.

We also performed an exploratory analysis to determine if the polysemy advantage was present throughout the age range within the child population (4½ to 7 years). We calculated a median split on age (median = 6.01 years) and entered each groups' data into the model with condition as the fixed effect (and random intercepts for subjects and items and a random slope and intercept for order). This model revealed a significant effect of condition (polysemy) (ß = 0. 25, t = 3.11, p = 0.01) in the younger half of participants. Using the same model used to predict performance in the older half, the model again revealed a significant effect of the polysemy condition (ß = 0.38, t = 5.89, p = 1.51e-05), and a model predicting performance across the two groups using condition and age group found no significant interaction (ß = -0.13262, t = -1.405, p = 0.166745), suggesting that the polysemy over ambiguity advantage is present throughout our child sample to a similar degree (random terms included an intercept for subjects and slopes and intercepts for items and block number).

Comparisons with chance

Final exploratory analyses include comparisons to chance. Interestingly, bootstrapped 95% confidence intervals show that children were above chance (0.25) in both conditions at timepoint 1 (ambiguity: M = 0.337, 95% C.I. = 0.278 - 0.393; polysemy: M = 0.643, 95% C.I. = 0.579 - 0.698), suggesting that children were to some extent also able to learn ambiguous words, consistent with past findings showing knowledge of ambiguous words at young ages (Backscheider & Gelman, 1995). Yet ambiguity was challenging for children,

given how close to chance their performance was in the ambiguity condition. Further, their retention of ambiguous words overlapped with chance after the week delay (ambiguity: M = 0.237, 95% CI [0.158 – 0.316]; polysemy: M = 0.657, 95% CI [0.556 – 0.731]), indicating that they required more input than our exposure provided in order to retain multiple, unrelated meanings. In adults, we found performance parallel to children at the first timepoint (ambiguity: M = 0.417, 95% CI = [0. 353- 0.476], polysemy: M = 0.722, 95% C.I. [0.659 - 0. 774]). However, their performance deteriorated after 7 days, resulting in a significant reduction in the polysemy condition across the timepoints, and worse performance overall at the second timepoint, where no polysemy over ambiguity learning advantage was evident. However, in a non-preregistered comparison to chance, we did find consistent above-chance performance in polysemy at the second timepoint for adults: M = 0.355, 95% CI [0.285 - 0.419], while their performance was not reliably above chance for ambiguity, as CIs overlapped with 0.25: M = 0.310, 95% CI [0.238 - 0.375].

Discussion. As predicted, both children and adults showed a significant polysemy over ambiguity advantage immediately following exposure, demonstrating that words with related meanings are easier to learn than words with unrelated meanings. The effect remained strong in children after a 1-week delay without re-exposure. Adult performance decayed after the week delay, and no polysemy over ambiguity advantage was evident (although performance only in polysemy condition remained above chance). There are two possible explanations for the difference between adults and children after delay without assuming that word learning is a critically different process in adults and children at this age. First, children were retested in person with the same experimenter and in the same location, while adults were re-recruited online. Because of this, children had additional cues to retrieve

their past experience with the task, and thus the exposure; we know contextual cues such as physical surroundings can have significant effects on lexical memory (Baddeley et al., 2009; Egstrom et al., 1972; Godden & Baddeley, 1975). Because adults were recruited online, it was impossible to ensure they were in a similar context at the second exposure.

Another, nonexclusive possibility is that adults' more densely-packed lexical and phonological networks impeded their ability to retrieve the novel word labels they had learned a week before. Recall that the task required participants to determine which of four meanings corresponded to the particular label they were prompted with. Accurate performance was possible by a process of elimination: if two novel meanings were recalled with different labels, the witnessed label must belong to one of the remaining options. This strategy would require specific memory of word labels for the other items each trial, however adults in particular have been shown to struggle in name- and word-retrieval tasks (Dell & Gordon, 2003; Ramscar et al., 2014). If their denser lexicons made the task more difficult, then a different task that primarily depends on memory of meanings rather than memory of label-to-meaning mappings may reveal lasting evidence for stronger performance on polysemous word meanings. In Experiment 2, we provided participants with a task that evaluated this possibility.

3.2 Experiment 2

Experiment 1 demonstrated that both children and adults have a stronger memory for polysemous word-to-meaning mappings when asked to discriminate them from differently-labeled distractors immediately after exposure, and for children, the polysemy over ambiguity advantage remained after a week delay. One concern about the task in Experiment

1 is that participants were only asked to assign a single meaning to each label on each trial, so that they may have relied on memory of one or two novel words to perform well. Another issue is that Experiment 1 required participants to construe differently-labeled targets as dissimilar, but it did not directly test whether participants were able to group together commonly-labeled polysemous targets as sharing the same label. Therefore, in Experiment 2, we ask participants to discriminate the targets from a set of foils that had been witnessed during the same exposure video.

Specifically, on each trial participants were presented with three items from a single meaning set along with five foils, and were prompted with a label. Because this task presents participants with all three targets simultaneously, we report a separate norming study with a different group of participants in order to rule out two possible alternative explanations for participants' success in learning multiple polysemous meanings. First, it is possible that a priori visual similarity between the polysemous items is sufficient to explain participants' successful identification of targets at test. It is also possible that the higher relative exposure to the prototype in the videos may have facilitated learning of the prototype in each set of meanings, but that participants simply used this representation to infer the other two targets on-the-fly, without learning the two additional meanings. In our norming task, we rule out both of these possibilities by giving participants the same exposure as in the main task without label information, and then asked them to select the three items which they think are likely to share a label. Since participants only saw items from differently-labeled sets in the prior experiment, examining these explanations was not possible in Experiment 1.

The task of Experiment 2 did not require memory for specific labels. It instead probed whether participants' memory for which items had been labeled to test whether identifica-

tion of the 3 target meanings of each word benefited from similarities between exemplars in the polysemy condition. This task then also allows us to address the possibility that adults' low performance at the second timepoint in Experiment 1 was due to the challenge of retaining new lexical items. If this led to a decay over time in adults' performance, we should see more robust retention at follow-up in Experiment 2.

Method Participants. The same 84 children and 84 adults from Experiment 1 were tested. As in Experiment 1, the same 37 children and 59 adults were tested again after a week, with no intervening exposure (again, children: N = 18 in polysemy, N = 19 in ambiguity; adults: N = 31 in polysemy, N = 28 in ambiguity).

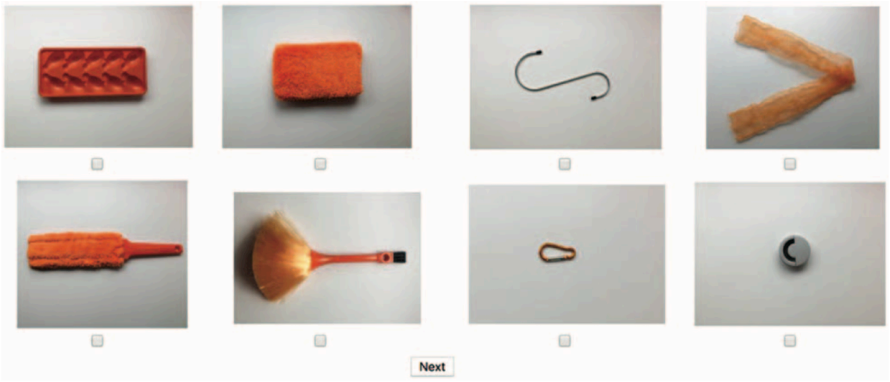
In a norming study, a separate group of 84 adult participants were recruited from Amazon's Mechanical Turk (preregistered). This group witnessed the identical visual exposure but without any verbal labels, ensuring that they would see the prototypical objects twice as in the main task, and were then presented with the identical test items used in the main analyses (see Test) with the following instructions: "Three of these objects share a label. Your job is to guess which ones."

Stimuli & Exposure. Adults and children took part in Experiment 2 immediately after Experiment 1 without any additional exposure. This order was used, rather than the reverse or a counterbalanced order, to minimize possible learning during the first task that would be relevant to the second task. Recall that Experiment 1 provided an indication that the targets on each trial had been labelled differently from one another since participants were asked to select the *kaisee/nona/gazer/veebo* from among a set of objects that had been witnessed with other labels. But it did not provide evidence of which three meanings belonged together with a label, which was required for accuracy in Experiment 2. Experiment 2, on

the other hand, provided information that could be used in Experiment 1, since in Experiment 2, only targets which matched the label in the prompt were displayed to participants, along with a set of foils that had not been labeled. Remaining concerns about Experiment 2 being unduly influenced by Experiment 1 were addressed by a separate sample of 92 adult participants who were tested only on Exp. 2. Those results confirmed the same pattern of data reported here. It remains possible that the testing at timepoint 2 was affected by the testing at timepoint 1, but this was unavoidable, and importantly, whatever advantage was gleaned at timepoint 1 for the polysemy condition was also available for the ambiguity condition.

Test. Participants performed a Sense Selection task in which they were shown three meanings of a word along with 5 of the foils from the corresponding exposure video, and were asked to "Pick three kaisees," (the 4 nonce labels counterbalanced across conditions, see Figure 3.6). The same subset of children described in Experiment 1 were re-tested again after the week-long delay, without additional exposure; they were simply shown the same test trials in the same order as they had been a week earlier.

Figure 3.6: Polysemy trial (one of four). Participants were prompted with a label and asked to select three target meanings. See the online article for the color version of this figure.



Results

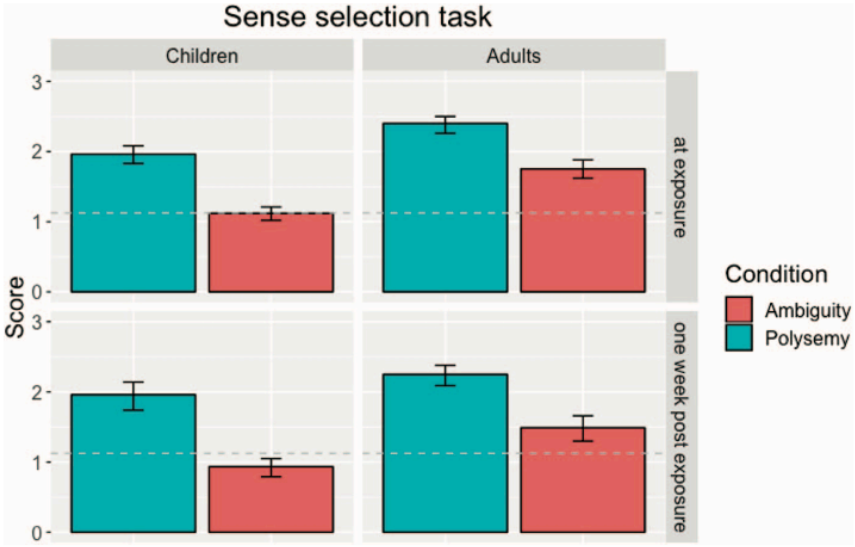
Adults. Figure 3.7 shows performance in the two conditions for both age groups, with chance at 1.125[†] The preregistered maximal converging multilevel model included condition as the fixed effect, and subject, order, and items as random intercepts with random slopes for item and subject. It revealed significantly better accuracy in the polysemy condition in comparison to the ambiguity condition (ß = 0.65, t = 4.26, p = 0.0005).

Children. The same model was preregistered for children, and the same maximal random structure converged for the child data. Children also performed significantly more accurately in the polysemy condition compared to the ambiguity condition (ß = 0.84, t = 8.76, p = 2.64e-13).

Norming. We calculated "item predictability" scores for each novel word's set of three meanings in each condition by averaging participants' performance in the Norming (no exposure) study on each trial, and used these scores as a measure of predictability on the basis of similarity and on the basis of inference from higher frequency of the prototype (sample size and analysis preregistered). Item predictability was included as fixed effects with random slopes and intercepts for subjects and order in both populations. The models for both adults and children still revealed a significant advantage for adults and children in the polysemy condition, even in the presence of item predictability (adults: ß = 0.59 , t = 2.82, p = 0.02), children: ß = 0.78, t = 4.83, p = 0.0006). In fact, visual predictability was not a significant predictor for either group: adults (ß = 0. 08, t = 0.18, p = 0.87), or children (ß = 0.15, t = 0.56 , p = 0.59). Thus, our prediction was confirmed that the polysemy over ambiguity advantage reflects a process of word learning and is not explained by the predictability

[†]Each trial offered participants to earn a score of 0, 1, 2, or 3 for selecting up to 3 correct targets out of 8, so, chance performance is calculated to be 1.125 out of 3.

Figure 3.7: Average accuracy in the Sense selection task for the ambiguity and polysemy groups. Child performance (4½ -7 years old) at exposure in upper left panel (n = 42 per condition), adults' performance at exposure in upper right panel (n = 42 per condition); 4½ -7-year-old children's performance one week later in bottom left panel (n = 18 in polysemy, n = 19 in ambiguity); adults' performance one week later in bottom right panel (n = 31 in polysemy, n = 28 in ambiguity). Error bars represent bootstrapped 95% Confidence Intervals.



of the items.

Longitudinal results. Once again, the exploratory analyses for timepoint 2 included timepoint and its interaction with condition as a fixed effect. The bottom panel of Figure 3.7 shows their accuracy averaged in each condition at the second timepoint in each group.

Children. The bottom left quadrant of Figure 3.7 shows children's scores on the second task one week after exposure. The model predicting performance of children at follow-up revealed a significant effect of the polysemy condition (ß = 1.02, t = 7.88, p = 1.63e-10) including random intercepts and slopes for participants and order, and random intercepts for item (this analysis required model comparison using between two models with equally complex random structure). To determine if there were changes in accuracy across time-

points, we again analyzed just those children who participated at both timepoints using a model with timepoint, condition, and their interaction as the fixed effects, and random intercepts and slopes for subjects, items, and order. The polysemy over ambiguity advantage held; in fact, there was no evidence of memory decay after the week-long delay. That is, there was a main effect of condition (polysemy) (ß = 0.96, t = 7.80, p = 2.05e-09), but no main effect of timepoint (T2) (ß = -0.17, t = -1.44, p = 0.16) and no significant interaction of condition (polysemy) and timepoint (T2) (ß = 0.06, t = 0.35, p = 0.73). Again, to ensure that these participants were representative of our main sample on this task as well as the preceding task, we entered results into a multilevel model to compare performance on this task at the first timepoint between the whole group and the subgroup which returned for the follow up, using group, condition, and their interaction as main effects with maximal random structure (random intercepts for subjects random slopes and intercepts for item and order). Reassuringly, this revealed no main effect of group (full group) (ß = -0.01, t = -0.11, p = 0.91) and no interaction of condition and group (ß = -0.07, t = -0.48, p = 0.64), but again revealed a main effect of the polysemy condition (ß = 0.84, t = 47, p =1.91e-10).

Adults. The bottom right quadrant of Figure 3.7 shows adults' scores on the second task one week after exposure. Consistent with both the adult data at the first timepoint and the child data at the first and second timepoints, the model (including random intercepts and slopes for subjects, items, and order) found a significant polysemy advantage at follow up (ß = 0.76, t = 3.17, p = 0.01).

We also analyzed across timepoints, taking just participants from both timepoints (n = 59), and again using a model with timepoint, condition, and their interaction as fixed effects and maximal converging random structure which included random intercepts and

slopes for subject, order, and item. This analysis revealed a significant main effect of the polysemy condition (ß = 0.77, t = 3.91, p = 0.0009), but no effect of timepoint (ß = -0.24, t = -1.24, p = 0.23) or interaction between timepoint and polysemy (ß = -0.01, t = -0.04, p = 0.97), suggesting that in this task, the polysemy advantage was present to the same extent across both timepoints.

To confirm that the returning adult participants were representative of the full adult sample in this task as well as the task reported in the first experiment, we entered results into a multilevel model to compare performance at exposure between the whole group with the returning subgroup, using group, condition, and their interaction as main effects (random intercepts and slopes for order, item, and subjects). This revealed no main effect of group (returning subgroup) (ß = -0.004, t = -0.05, p = 0.96) and no interaction of condition (polysemy) and group (ß = 0.03, t = 0.24, p = 0.81), but a main effect of the polysemy condition (ß = 0.65, t = 4.04, p = 0.0004).

Age comparisons. To determine whether there was an effect of age on performance at the first timepoint on the Sense Selection task, we fit a model with age group (adults vs. kids) and condition as fixed effects (random slopes and intercepts for subjects, item, and order). This exploratory model revealed a significant effect of condition (polysemy) (ß = 0.65, t = 4.67, p = 0.0004) and age group (children) (ß = -0.63, t = -4.52, p = 0.002141), and no significant interaction of condition (polysemy) and age group (children) (ß = 0.19, t = 1.05, p = 0.31). Adults' better performance suggests that unlike what was found for the Label Matching task of Experiment 1, the benefit of polysemy over ambiguity in learner's ability to discriminate targets from foils in the Sense Selection task may continue to develop with age. At the second timepoint, we also compared children and adults using a model

with condition, age group, and their interaction as fixed effects (random effects and intercepts for subjects, items, and order), and found a significant effect of polysemy (ß = 0.76, t = 4.37, p = 0.001) and of age group (children) (ß = -0.56, t = -3.07, p = 0.006), but no significant interaction (ß = 0.27, t = 1.07, p = 0.30). So, while the polysemy advantage was present across both groups, and children didn't do as well as adults overall, there was no evidence that the polysemy advantage was attenuated in children in particular.

In order to investigate possible age effects further, in another non-preregistered analysis, we entered child data with a median split by age (median = 6.01 years) to determine if the polysemy advantage was present throughout the age range in the sample. First, we entered each half of the data into a multilevel model with condition as the fixed effect (random intercepts for subjects and items and a random slope and intercept for order). This model once again revealed a significant effect of condition (polysemy) (ß = 0.73, t = 2.95, p = 0.01) in the younger half of participants, and the same model found that the polysemy condition again predicted better performance (ß = 0.97, t = 6.68, p = 1.38e-06). A model predicting performance across the two groups using condition and age group (random intercepts and slopes for items and order and random intercepts for subjects) found no significant interaction of age group within the child sample (older half) and condition (polysemy) (ß = -0.24, t = -0.86, p = 0.40), again finding a main effect of condition (polysemy) (ß = 0.97, t = 6.40, p = 1.23e-07), and no main effect of age group (ß = -0.14, t = -0.72, p = 0.48).

Comparisons to chance. Bootstrapped 95% confidence intervals were used to compare to chance performance of selecting between 0 and 3 correct targets out of a possible 8 items. These exploratory analyses showed that children were reliably above chance in the polysemy condition at timepoint 1 (polysemy: M = 1.96, 95% C.I. [1.83 - 2.08], but not in

the ambiguity condition as the bootstrapped 95% CI overlap with chance [1.03 -1.21]. Furthermore, while children's performance after a week delay remained above chance in polysemy (M = 1.96), 95% C.I. [1.74 - 2.15], performance in ambiguity fell reliably below (M = 0.93), 95% C.I. [0.79 - 1.07]. Such low performance in children at follow-up in Experiment 2 may be considered surprising, as adults seemed to do comparatively better in Experiment 2 at follow-up as compared to Experiment 1. At timepoint 1 in Experiment 2, adults were also reliably above chance (ambiguity: M = 1.75, 95% C.I. [1.62 – 1.88]; polysemy: M = 2.40, 95% C.I. [2.26 – 2.50]), but unlike in Experiment 1, adult performance was reliably above chance in both conditions by the second timepoint (ambiguity: M = 1.49, 95% CI = 1.30 - 1.66, polysemy: M = 2.25, 95% CI = 2.09 - 2.38).

Discussion. Children and adults both showed a significant polysemy over ambiguity advantage and the younger half of children (4 ½ - 6) performed as well and showed as strong of a polysemy advantage as the older children. Both children and adults retained the polysemy advantage after a 1-week delay without re-exposure. Norming results confirm that this robust polysemy advantage cannot be attributed to visual similarity among items, as including visual similarity in the model had no significant effect for either children or adults. Adults outperformed children at both timepoints, unlike in Exp. 1, which is consistent with the idea that adults' poorer performance after the delay in Exp. 1 was likely due to the challenge of retrieving new (low-frequency) words from their denser lexicons.

General Discussion

The present experiments investigated children and adults' ability to learn 4 novel words which were assigned three distinct meanings apiece. We manipulated whether the three meanings of each word were related to one another (polysemy condition) or not (ambi-

guity condition). The meanings of each word in the polysemy condition were designed so that they shared no single distinguishing feature but instead varied along two different dimensions, reducing the possibility that participants could rely on a simple definition or a single extension rule in order to distinguish the three target meanings for each word from foils. Recall that novel labels and novel meanings were used in order to mitigate the effects of previous experience. No feedback was provided, and available participants were tested immediately after exposure and again after a one-week delay.

In each of two tasks, we consistently found a polysemy over ambiguity advantage in learning the novel words immediately after minimal exposure. That is, children and adults who were exposed to 4 novel polysemous words outperformed those who were exposed to 4 ambiguous words in their ability to map novel labels to the appropriate target meaning, distinguishing that meaning from otherwise-labeled distractors (Experiment 1). Participants who had witnessed four polysemous words also performed better at correctly identifying and distinguishing multiple target meanings for each novel word from foils (Experiment 2). Since our interest is in vocabulary learning, which is consolidated in long term memory, we retested participants after a week-long delay without providing additional exposure.

After confirming that the participants who took part in the testing after a week did not differ significantly in age or performance from the full group, we performed the identical analyses for the data at the second time point. Remarkably, children's performance in the polysemy condition after a week remained well above their performance in the ambiguity condition in the label mapping task (Exp.1) and in fact showed no decay after a week in the sense selection task (Exp. 2). Adults' performance fell to chance in the label mapping

task after the weeklong delay, possibly because adults' larger lexicons made the retrieval of target words from long-term memory more demanding (Dell & Gordon, 2003; Ramscar et al., 2014). In the second task, which required participants to discriminate between commonly-grouped meanings and foils and did not require lexical access, adults showed the same polysemy > ambiguity advantage evident in children (Experiment 2).

The present findings are consistent with work on non-linguistic category learning insofar as meanings that shared more overlapping features were easier to categorize together than those that shared fewer distinguishing features. This suggests, perhaps not surprisingly, that word-learning may well be a special type of category learning, as long as "category" is not assumed to require a single vague or summary representation, but is instead understood to allow a range of subtypes with their own unique properties. Word learning is special insofar as learners need to use the labels as a cue to multidimensional category membership, and must learn the arbitrary mappings between exemplars and labels without receiving explicit instruction nor corrective feedback. The current findings are also consistent with prior work which suggests that ambiguous meanings of a word compete with one another during comprehension but polysemous meanings do not. If polysemous meanings are represented as partially overlapping, distributed representations (Rodd et al., 2012), polysemous meanings of a word may reinforce one another rather than compete with one another as is assumed by current models of word learning.

Particularly because the polysemy we investigate is irregular, we do not assume that children necessarily expect or predict that the word label will extend in any particular way. The current exposure provided positive evidence that the distinct meanings were labelled by the same word, which appears to provide an advantage during integration and long-term

consolidation. Earlier work had emphasized that unrelated meanings of a word label were easier to learn when each meaning was used as a distinct syntactic category or was from a wholly different semantic field (such as labeling a novel animal a glass). For example, consistent with the current findings, Casenhiser (2005) found that children were reluctant to assign a familiar noun like *dog*, another, unrelated noun meaning such as "monster." At the same time, children were willing to reuse a label from a different grammatical category (a *did* could be a monster). That is, ambiguous meanings are less difficult to learn when they apply in easily distinguishable contexts rather than in contexts that have the potential to overlap. In the current work, all novel words referred to artifacts and were used in nominal contexts, so the contexts provided were underdetermined and overlapping.

Is it possible that participants in the polysemy condition did not learn multiple meanings, but instead only learned the prototypical meaning and then generalized that label to the other two meanings on the fly during the tasks? If participants somehow only learned the prototype, they could be expected to perform better in the polysemy condition than in the homonymy condition, since by design, the additional polysemous meanings were related to the prototype while the homonymous meanings were not. Yet, an explanation of why and how learners might only learn the prototype during exposure is required to make this route to better performance in the polysemy condition feasible. In natural languages, learning only one prototype per form is not sufficient, since extensions of the same prototypical meaning vary across different languages. Moreover, the results of the norming study show that exposing learners to two instances of the prototype instead of one does not predict performance on the task in either children or adults. We suggest that implicit recognition and comparison of the 3 labelled meanings results in their overlapping features be-

ing strengthened. That is, implicit recognition and comparison of the 3 labelled meanings would allow a prototype to emerge, but critically, this perspective involves the recognition and comparison of multiple meanings.

If polysemy were rare in natural language, the question of how words with multiple meanings are learned might be reasonably put aside until the learning of words with single senses were better understood. However, polysemy is ubiquitous, particularly in the case of highly frequent and early learned words, and it commonly includes distinct meanings from the same grammatical category and semantic field (as is the case for the various senses of cap and other examples in the introduction). The relative lack of previous work on children's learning of conventional, distinct but related meanings of words is surprising insofar as there has been a great deal of work emphasizing a finding that is the inverse of the multiple-meaning phenomenon reported here. In particular, learners generally disprefer assigning a second label to a concept (Clark, 1987; Markman & Wachtel, 1988; Merriman et al., 1989). That is, learning synonymous labels is challenging and true synonymy is rare in languages. Near synonyms are almost always distinguishable in one way or another: e.g., by dialect (*pop* vs. *soda*), register (*buy* vs. *purchase*), or attitude (*thrifty* vs. *stingy*). Rather than assigning multiple labels to the same concept (synonymy), both polysemy and ambiguity involve assigning multiple concepts to the same label.

Production-based accounts have offered an explanation as to why polysemy and ambiguity are common across the world's languages by emphasizing the advantage of reusing words that are easy to access and produce (Harmon & Kapatsinski, 2017; Piantadosi et al., 2012). But it is important to distinguish polysemy from ambiguity, even though the distinction is undoubtedly gradient, because polysemy is so much more common than am-

biguity (Dautriche, 2015b; Navigli & Ponzetto, 2012a; Rodd et al., 2002b). The current work provides the first evidence that we know of demonstrating that conventional ("irregular") polysemous senses are easier to comprehend and retain than ambiguous senses.

We found little evidence of a developmental shift in the ability to learn novel polysemous words. Older children were not significantly more accurate in either task than younger children. Children in the label matching task performed as well as adults initially, and outperformed adults after the delay. Only in Experiment 2, which required participants to recall which meaning(s) were related, given a label, did adults significantly outperform children at both timepoints. Future work should test younger children to determine whether a polysemy over ambiguity advantage is evident in children younger than 4½.

The current results raise new questions. Given that the current polysemy over ambiguity advantage was equally evident in younger and older children, we can ask, do children demonstrate the advantage at the very outset of vocabulary learning or do they need to learn to learn polysemous meanings? Since novel words can be learned to apply to related meanings after such limited and potentially confusing exposure, how it is that learners constrain meanings to apply to just the range that is specific to that word in the given language? This question is addressed in a subsequent chapter. The results also raise questions about populations that may have more difficulty recognizing relatedness among meanings. For example, individuals on the Autism spectrum have been claimed to hyper-focus on distinctions at the expense of recognizing relationships. Does this imply that individuals on the Autism spectrum lack the same polysemy over ambiguity advantage evident in neurotypicals? This question is addressed in Chapter 5. What underlies the relationship between meanings? Is it still possible that predictable or productive rules or patterns can account for

people's expectations about polysemy, or does general semantic similarity seem more likely? This is considered in the fifth and final empirical chapter of the dissertation. Finally, in Chapter 7 I outline a counter-proposal to the competitive accounts of word learning which the present chapter casts doubt on.

3.3 Chapter summary

Many words are associated with more than a single meaning, and quite often, meanings of a word are semantically related to one another. This is especially true in the case of frequent words, many of which are learned by young children. While certain meaning extensions can be predicted by very general rules and therefore may be created on-the-fly, other meaning extensions need to be learned on the basis of experience because they are idiosyncratic to individual words. We predicted that the existence of semantic relationships among meanings of a word (polysemy) would facilitate the learning of that word when compared with learning a word that was associated with the same number of unrelated senses (ambiguity). Experiments exposed adults and 4½-7-year-old children to 4 novel words, with 3 meanings each, during 2 minutes of exposure videos in which 20 foils were interspersed. The novel words were created using novel objects and novel labels in order to reduce interference from familiar labels or meanings. Two preregistered experiments confirmed our hypothesis, finding better accuracy for both age groups in the polysemy condition on both tasks: they were better able to identify which meaning was assigned a given label, a task that required distinguishing one of the label's meanings from meanings associated with other labels (Exp. 1); and they were more accurate at the selection of multiple meanings for each word from among also-witnessed but unlabeled foils (Exp. 2).

Remarkably, after a full week's delay, the subset of children who were available to be retested again showed the same polysemy over ambiguity advantage, despite having had no additional exposure. Moreover, the younger half of children performed as well as the older half, so that no clear developmental changes in the polysemy over ambiguity advantage were evident. Adult performance after the delay was mixed: in Exp. 1, only the polysemy condition remained above chance, but direct evidence of a polysemy over ambiguity advantage was absent, due to lower performance overall. On the other hand, adults did display a lasting polysemy over ambiguity advantage in Exp. 2, which did not require them to access and compare the meanings of multiple word labels. The recognition that polysemous meanings are ubiquitous and relatively easy to learn implies that the working assumption, commonly made in modeling work on human word-learning that words map uniquely onto a single meaning or that multiple potential meanings of a word are always in competition with one another, should be retired.

4

Toddlers recognize multiple meanings—can word learning theories keep up?

Learning the meaning of a word is recognized to be a challenging task (Quine, 1960). Yet arguably, the majority of word learning theories still underestimate the com-

plexity of the problem. In particular, most work has overlooked or explicitly set aside the fact that up to 80% of frequently used words in English have more than a single meaning and can refer to multiple, related meanings, a phenomenon known as polysemy. For example, *collar* can refer to a dog collar or shirt collar, and *cap* can refer to a baseball cap, pen cap, or bottle cap (Lakoff, 1987; Rodd et al., 2002b; Zipf, 1945a). Yet many current models of word learning rely on competition between meanings: learning a candidate meaning may either directly delete (Trueswell et al., 2013b) or weaken (Stevens et al., 2017a; Yurovsky et al., 2013) other meanings.

According to one early influential theory called Propose-but-Verify (PbV), the word learning process consists of hypothesizing a single meaning which is either retained in subsequent encounters if confirmed (i.e., if the meaning is present or inferrable when the word is heard), or deleted entirely if disconfirmed. Disconfirmation occurs when the word is heard in a context where the previously hypothesized meaning is not present or intended; for instance, an hypothesized meaning "bottle cap" for the label *cap* will be deleted when *cap* is subsequently used in the absence of a bottle cap; Gleitman et al. 2005; Medina et al. 2011; Trueswell et al. 2013a). When a meaning is disconfirmed in this way, PbV predicts that an entirely new, independent meaning is hypothesized. While intended to explain how learners manage to converge on correct meanings given an underdetermined environment (Quine, 1960), this proposal predicts that it is not possible to learn more than one meaning of a word. That is, whenever a PbV learner encounters a second meaning of any word, the first meaning is presumed to be eliminated.

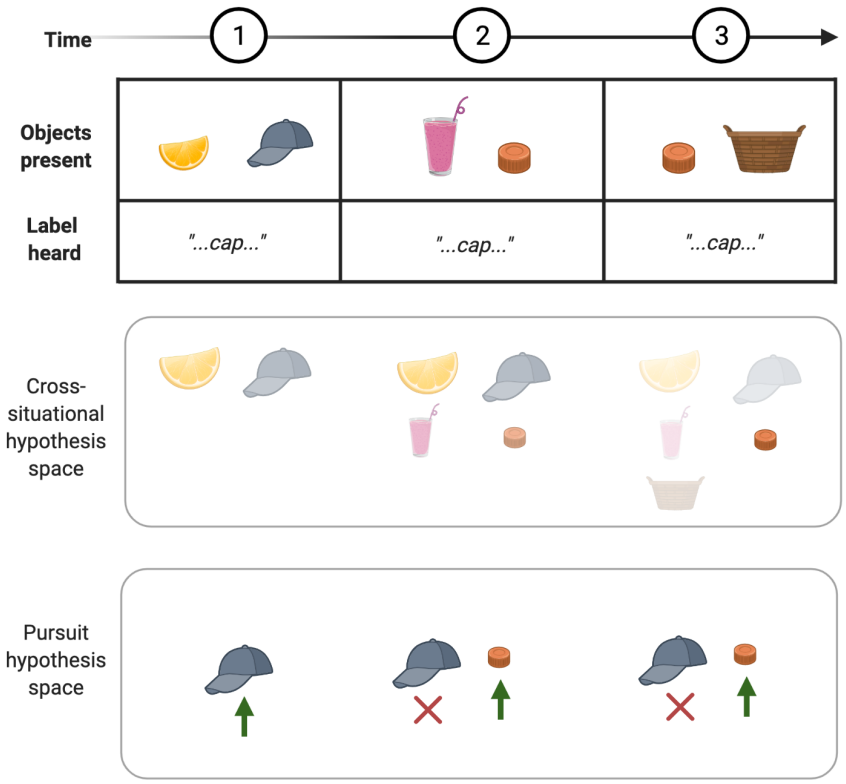
Other proposals allow for the possibility of learning more than one correct meaning for a word, in principle. **Cross-situational word learning** proposes that learners encode all

possible referents for a wordform over many successive instances, taking a union over the learned associations to determine a word's meaning (Yu & Smith, 2007). In principle, this kind of mechanism could allow for more than one meaning to become strongly associated with a word. Critically, however, successful cross-situational word learning models have relied on competitive mechanisms between candidate meanings across scenes in order to learn successfully, meaning that evidence for one meaning (such as the "bottle cap" meaning of *cap*) results in weakening of other candidate meanings (such as the "baseball cap" meaning) (Fazly et al., 2010; Xu & Tenenbaum, 2007; Yu, 2008; **?**). It is possible that exposure to the more common meaning, which is by nature more frequent, results in repeated suppression, leading to tenuous or nonexistent associations to the additional meanings. However, a lack of research testing such models on naturalistic, polysemous language input stands in the way of answering such questions.

More recently, a **Pursuit** model of word leaning has been proposed as an update to Propose-but-Verify (Stevens et al., 2016). Rather than automatically eliminating a previously hypothesized meaning as in PbV, this proposal allows for the disconfirmed hypothesis to be incrementally suppressed (e.g. weakening the hypothesis "bottle cap" after encountering the label *cap* in its absence). Though this mechanism allows for an additional meaning to be stored, the model faces the same challenges as cross-situational word learning when it comes to low-frequency meanings: encounters with a far more common meaning would lead potential low-frequency meanings to become tenuous or nonexistent through repeated suppression. In summary, though both Pursuit and cross-situational word learning hypothetically allow for the possibility of multiple meanings, the competitive mechanisms which these models rely on are likely to inhibit successful learning of additional,

correct word meanings (see Figure 4.1). And, though competition between meanings may be an efficient way identify the most common meaning, it is unclear how this approach can simultaneously ensure the viability of additional, relatively lower-frequency meanings.

Figure 4.1: Schematic of learning in both Cross-situational and Pursuit models. Across three instances, additional meanings become less likely and may become entirely eliminated.



While it is not completely clear how Cross-situational and Pursuit models would fare with polysemous words, recent research has suggested that human word learners are adept at tracking multiple meanings. Children as young as 3 years of age are able to learn novel, related meanings if they are related in a systematic way, *e.g.*, *the flour* vs. *some flour* (Srini-

vasan et al., 2019). Children as young as 4 years as well as adults learn and retain polysemous meanings which relate in unpredictable, convention-based ways, as is common in natural language, e.g. *baseball cap, lens cap* (Floyd & Goldberg, 2020). While these studies show that word learners succeed in tracking and retaining multiple, related meanings, it is not clear whether this is a foundational component of word learning. Indeed, it may be that early word learning requires the idealized situation commonly assumed by word learning theories, in which each word refers to a single meaning. That is, while children ultimately need to learn polysemy, it is possible that only after years of single-meaning learning do children acquire the metalinguistic skills to support polysemy learning. One recent study showed that toddlers' ability to learn artificial homophones (multiple, unrelated meanings) depended crucially on them being instances of distinct syntactic category (Dautriche et al., 2018). We contribute to this area by testing whether the semantic relationships in naturalistic, polysemous words allows toddlers to track more than one meaning from their linguistic environment.

To determine if early word learning mechanisms can support learning multiple meanings for individual words, we test young children, while vocabulary and metalinguistic skills remain limited, and use naturalistic, related meanings which are plausibly present in learners' environments. Therefore, using eye-tracking, the present study investigates whether or not 2 ½-year-old children are able to recognize polysemous meanings in English that refer to distinct concepts (e.g., baseball *cap* and pen *cap*).

It is important to recognize that some word meanings can be captured by a single vague or underspecified meaning rather than multiple distinct meanings (Falkum, 2015; Nunberg, 1979a; Tuggy, 1993). Therefore, success in recognizing more than one polysemous

meaning for a word such as *cap* may not actually reflect that participants are tracking more than one meaning, as this recognition can be equally predicted without positing two different meanings by simply underspecifying a single meaning that generalizes over both, i.e., "anything that covers" (Langacker, 1987). To address this concern, we include trials which test monolingual English toddlers on additional plausible meaning extensions which are present in another language (Spanish), but not in English (i.e. testing whether participants look to "lid" when prompted with *cap*).if toddlers rely on a single, underspecified word meaning to guess or infer target meanings during comprehension, performance should be identical across trial types (e.g. equal success identifying "bottle cap" and "lid" when prompted with cap, despite the fact that "lid" is not an attested meaning for cap in participants' English-speaking linguistic environments). On the other hand, if toddlers' performance on English meanings (e.g. "bottle cap" and "baseball cap" for *cap*) truly reflects recognition of more than one distinct meaning, they should not perform as well on novel extensions (e.g. should look less to "lid" than "bottle cap" or "baseball cap" when prompted with *cap*).

In this study, we hypothesized that children would successfully identify polysemous words on both trial types, but would perform reliably better on English meaning trials ("bottle cap", "baseball cap") than on Novel extension trials ("lid") in response to the same prompt (*Look at the cap*). This would imply that toddlers are able to guess new meanings of familiar words in real time, but critically, it would imply they track and represent multiple meanings for words in ways that are consistent with their English input. Finally, in order to determine whether existing proposals of word learning can account for knowledge of more than one meaning per word, we implemented two models of word learning with nat-

uralistic, child-directed polysemous input. Specifically, we selected Cross-situational word learning and Pursuit models because their learning mechanisms do not in principle preclude the possibility of learning more than one meaning, and because they have previously been shown to learn single-meaning words successfully from naturalistic input.

4.1 Experiment 3

Monolingual English-speaking children were tested on two meanings of each of six English words. We also included a third novel meaning for each of the six words, choosing extended meanings that are used in another language (Spanish) but not in English. On trials in the English meanings condition, children were exposed to images of conventional word meanings, such as "bottle cap" and "baseball cap", opposite a distractor (which was a target item on a separate trial) and prompted to look to the target meaning using the same prompts for each pair of meanings (e.g. *Look at the cap!*). On trials in the Novel extension condition, children heard the same English labels and prompts, but targets were unattested yet plausible polysemous meanings (e.g., "lid"). We compared children's performance on the English meaning trials (e.g., "baseball cap", "bottle cap") to chance and to their performance on Novel extension trials which involved unfamiliar but potentially inferrable meanings (such as *cap* referring to a lid).

Method

Participants. Forty 24-36-month-olds (M = 31.2 months, SD = 2.6 months; 21 females) with typical hearing and vision development participated in an eye-tracking experiment, two of whom (n=2) were excluded for excessive trackloss (n = 2) for a total sample of 38 children (see Results section for exclusion criteria). An additional 10 toddlers were re-

cruited but not included from the study due to inability to complete calibration (n = 8), parental interference (n = 1), and participant noncompliance (n = 1). All children were monolingual and exclusively exposed to English. Participants were tested in both conditions (English meanings and Novel extension).

Stimuli. Three criteria were used to select the six target words: their imageability, their presence in the CHILDES corpus for children under age 3 (MacWhinney, 2000), as well as whether their translations in another language (Spanish) were associated with a polysemous extension not shared by English (see Table 4.1 ; *e.g.* a single label is used in dialects of Spanish for at least one of the English meanings and the Novel meaning). This allowed us to ensure that the Novel meanings were plausible extensions for the words.

Table 4.1: Target polysemous words and meanings. In the Novel extension column, the corresponding Spanish word is in italics.

Word label	English meanings	Novel meaning
cap	bottle cap, baseball cap	lid (*tapa*)
sheet	sheet of paper, bed sheet	leaf (*hoja*)
glasses	eyeglasses, drinking glasses	goggles (*gafas*)
collar	shirt collar, dog collar	necklace (*collar*)
horn	animal horn, musical horn	croissant (*cuerno*)
balloon	party balloon, hot air balloon	globe (*globo*)

Procedure. All toddlers sat on their caregivers' laps (with the exception of one child who sat alone in the chair), approximately 50 cm from the screen in a sound-attenuated

room. The caregivers wore an eye mask which blocked them from seeing what was on screen, preventing a potential source of bias, such as leaning or pointing. The eyetracker was calibrated for each child using a three-point calibration on an EyeLink 1000 Plus. The experiment, which lasted approximately 3 minutes, was created using Experiment Builder software (SR Research, Mississauga, Ontario, Canada) and controlled from a Mac host computer. Toddlers were shown stimuli on a 17-inch monitor while the eye tracker sampled the location of their eye fixations at 500 Hz. On each of 18 test trials, children were prompted with simple English sentences, e.g. *Look at the cap!* while two images were displayed on the screen. There were also four filler trials, which displayed a rainbow circle or pinwheel accompanied by a chime sound.

Word Familiarity Questionnaire. To check whether toddlers had been exposed to both conventional meanings of each English word, we asked parents about their child's input. Parents were given a questionnaire which was administered on a laptop after their child completed the study. The questionnaire asked parents to rate how familiar their child was with each of the meanings (e.g. "How familiar is your child with the word cap as in baseball cap?") on a 5-point scale between "not at all, has never heard it" to "definitely familiar, hears it often."

Results

For all trials, we identified looks to target and distractor images within the 500x500 pixel size of each image, and coded track loss as looks outside of these regions. We excluded trials with greater than 75% track loss. This resulted in the loss of 17.9% of all trials, including 2 subjects entirely, as they had greater than 75% track loss on every trial. Using RStudio, we averaged samples across 100-ms time bins and calculated the proportion looking to target

(i.e., the number of samples which were looks to target divided by the sum of samples to target and distractor). We calculated toddlers' proportion of looks to target for each subject in each trial type (English meaning and Novel extension), collapsing across the window of 300-1800ms following noun onset, as in previous research (Fernald et al., 2013; Lukyanenko & Fisher, 2016).

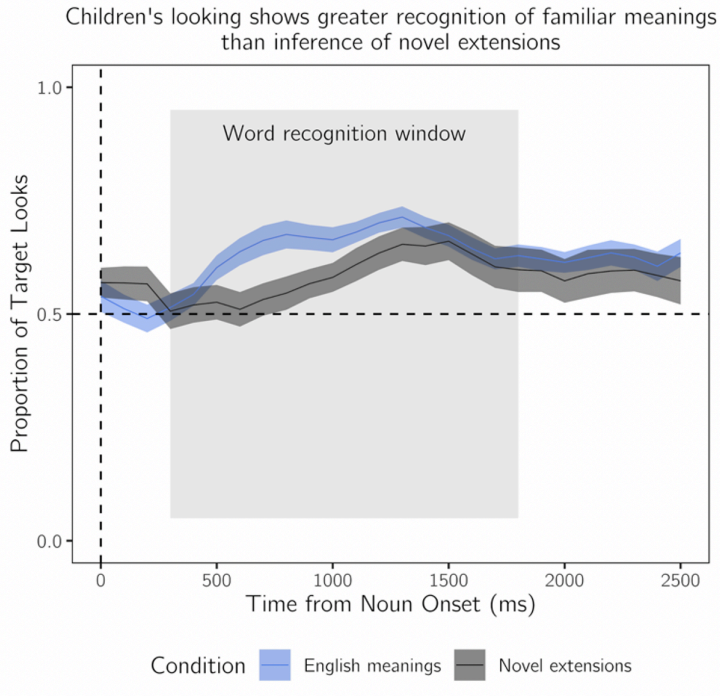
We predicted that while performance in both English meaning and Novel extension trials would be above chance, performance on English meaning trials should be significantly higher than on Novel extension trials. This was based on our hypothesis that children can learn multiple meanings for labels which are attested in their language, allowing them to better recognize these attested multiple meanings, as compared to the unattested, novel extensions. At the same time, predicted that they would also be able to infer novel extended meanings when faced with a choice between a plausible extension vs. an unrelated distractor.

To first evaluate overall performance, we entered data into simple one-sample t-tests against chance for each condition. Toddlers' mean proportion of looks to target on English meaning trials during the window of interest was significantly above chance (0.5), $t(37) = 7.79$, $p < 0.001$. Toddlers were also able to reliably infer the correct target in the Novel extension trials ($t(37) = 3.30$, $p<0.001$). The final prediction was that, if children at this age are able to learn multiple meanings from their linguistic environment, as opposed to a single vague or underspecific meaning, they should perform significantly better when required to recognize English polysemous meanings than meaning extensions which they had never heard before (Novel extension trials).

To test this prediction, we calculated average proportion of looks to target in each of

the two conditions (English meanings and Novel extensions) for each subject and entered it into a paired *t*-test, which revealed that children looked significantly more to target for English meanings than for Novel extensions ($t(37)=2.45$), $p < 0.05$). We confirmed these results with a maximal converging mixed effects model, selected by the buildmer package (Voeten, 2020), with random slopes and intercepts for subjects. This model revealed that performance in the Novel extensions condition was significantly lower than in the English meanings condition ($\beta = -0.06$, $t = -2.53$, $p = 0.01$), indicating that our participants could already recognize multiple meanings from prior exposure to English polysemous words, beyond what simply guessing a secondary meaning based on an underspecified representation would allow, as in the Novel extensions condition (see Figure 4.2).

Figure 4.2: Proportion of looking to target following noun onset (shown at 0ms) in toddlers (n = 38) ages 24-36 months on English meaning and Novel extension trials. The window of interest is 300 - 1,800ms following noun onset.

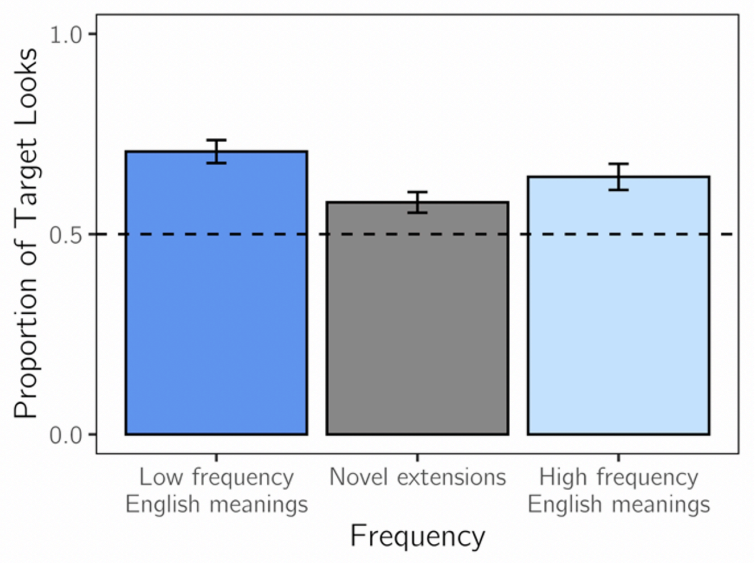


One possible concern is whether toddlers actually recognized both meanings of each polysemous noun, or whether instead they used familiarity with one meaning (or a general/underspecified meaning) to succeed in the Novel extension (unfamiliar meaning) condition, or even to use a single, higher-frequency English meaning to succeed on an unfamiliar English meaning. To rule these possibilities out, we used parental report to split the English meanings into two groups for each child: higher and lower familiarity. We then compared both to performance on Novel Meaning trials (Figure 4.3). We excluded trials for which the caregiver estimated equal familiarity for both meanings of the word (34%), as those were not informative about the relative familiarity of the two meanings. In order to simultaneously compare all three trial types, we again entered data (again summarized into 100ms time bins per subject per trial within the word recognition time window) into a multilevel model with group (higher-familiarity English meaning, lower-familiarity English meaning, and Novel extension) as the fixed effect along with maximal converging random structure (random intercepts and slopes for subjects). We again found that performance on English meaning trials was higher than on Novel extension trials, and though performance on the higher-familiarity trials vs. Novel did not reach significance, it was marginal and consistent with the predicted direction (lower familiarity: $\beta = 0.13$, t = 4.29, p<0.0001; higher familiarity: $\beta = 0.07$, t = 1.64, p =0.09). Multilevel model results with the reference levels reordered confirmed that there was no difference between the higher- and lower-frequency meanings ($\beta = 0.02$, $t = 1.29$, $p = 0.2$).

Critically, there is another possible explanation for lower performance on Novel Meaning trials: that the Novel extension meanings (e.g. "lid", "croissant") were already known to children under English labels (e.g. lid, croissant), and children could have been slower to

look at the lid when prompted with *cap* because the "lid" meaning was already preempted by another label in their vocabulary. If participants performed worse on Novel meaning trials because they faced interference from these labels, greater familiarity with the label (earlier age of acquisition for that label) should be inversely correlated with their accuracy in looking to the target on those trials. We therefore analyzed our data for this possibility, using data from age of acquisition norms (Kuperman et al., 2012). We found that, not only that the age of acquisition (AoA) for these labels is far later than the age of our participants (see Figure 4.4), but also that the relative AoA of the competing labels does not predict lower performance on these trials: a linear model predicting accurate looking to target on Novel Meaning trials showed no significant negative effect of age of acquisition of the true English labels on performance in these trials ($\beta = 0.01$, $t = 0.73$, $p = 0.47$).

Figure 4.3: Performance summarized over recognition window for Higher-frequency English meanings, Lower-frequency English meanings, and Novel extensions.



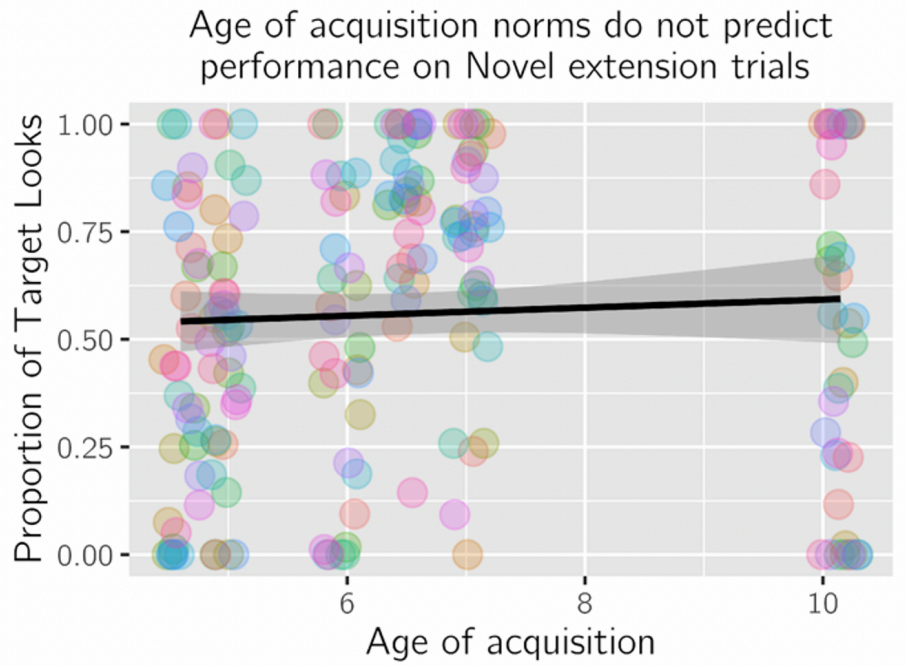
4.2 Experiment 3 modeling study

Results from the prior Experiment show that toddlers have learned more than a single correct meaning for a set of polysemous words. Recall that current theories of word learning have not been quantitatively evaluated to determine if they can account for this kind of multi-meaning learning. In fact, very few proposals which we know of explicitly allow for the hypothetical possibility that a model could track more than a single meaning for a word. Here, we present the first work to test two current models (Cross-situational and Pursuit) for their ability to learn more than one meaning for words from naturalistic utterances present in children's linguistic environments (CHILDES). While other models of early word learning have been proposed (see McMurray et al. 2012 for a review), we selected the two which represent central approaches about word learning: Cross-situational being a global model which aggregates associations across all learning instances, and Pursuit, a local/hypothesis-testing model, which learns one hypothesized meaning per encounter. These two models showed the greatest success in training based on naturalistic data in prior work which had tested the learning of single-meaning words (Stevens et al., 2016), and have an in-principle capacity to represent single- and multi-meaning words within the same model, unlike the models proposed by McMurray et al. which additionally require up to 200,000 learning epochs per word, outpacing the naturalistic data available. Finally, both Cross-situational and Pursuit models can be tested on identical input, which allows us to control for the task structure.

Methods

To determine how these models fare in learning multiple meanings, we compare how

Figure 4.4: Each dot represents average performance of one child on each of 6 Novel extension trials. If familiarity with a competing label created competition, driving down performance on Novel extension trials, then the later the word is acquired, the lower performance should be in our toddler participants. However, this pattern was not borne out in the data, suggesting that possible familiarity with the correct English label in Novel extension trials is not responsible for the lower performance. Instead, we suggest that the difference between the English trials and the Novel trials was simply that the children had already become familiar with the conventional meanings while they had not been exposed to the novel meaning.



both Cross-situational and Pursuit models fare in learning both single-meaning and multi-meaning words. To do this, we provide the models with CHILDES input containing the polysemous words used in our eye tracking experiment. We use CHILDES input from all corpora which contained video. In Stevens et al., additional versions of Cross-situational and hypothesis-testing models were also tested, but are not included here because they are either acknowledged to be unable to represent more than a single meaning of a word (Propose but Verify; Trueswell et al. 2013a) or are theoretically similar to the existing cross-situational model, but showed worse performance on the task in Stevens et al. (Fazly et al.,

2010).

Input to the model. Following the Stevens et al. coding scheme, we coded verbal interactions and accompanying concrete noun meanings which were visible to the learner according to video input from CHILDES. We included all utterances spoken to children between the ages of 0 to 36 months (the maximum age of participants in the preceding Experiment) which contained the target words in our experiment as well as their plural counterparts where possible, following the procedure in Stevens et al., 2016. In order to ensure the models did not fail to learn polysemy solely because of limited input, we expanded the input to include all transcripts from CHILDES which contained accompanying video, rather than just a single corpus as in the prior comparison. *Collar*[*] was excluded, as the video input did not clearly contain more than one meaning, and so were plural variants for which only one learnable meaning was present. This resulted in a total of 7 target polysemopus words, as were an additional 14 single-meaning words which appeared alongside the polysemous target words. Following Stevens et al.'s criteria, these were included as target meanings as well. In total, we included 284 transcribed utterances and accompanying visual input (coded) for the 21 target words (an average of 13.5 utterances per word) [†].

Model implementation. We ran implementations of the cross-situational model (Alishahi et al., 2012) and the pursuit hypothesis-testing model (Stevens et al., 2016) exactly as they had been implemented in a model comparison on single-meaning input (Stevens et al., 2016). All parameters were optimized with respect to the same overall score for both

[*]Though both meanings of *collar* appear elsewhere in CHILDES transcripts, we were restricted to the video-containing subset of CHILDES as candidate meanings had to be coded from visual input following Stevens et al.

[†]This ensured a close match to Stevens et al., which had included 496 utterances and accompanying visual input for 34 words (14.5 utterances per word).

models, following the past implementation.

Results

To determine how successfully each model performed, we first calculated what percentage of single-meaning words were accurately learned (see Figure 4.5). Here, we conceptually replicate (Stevens et al., 2016), in which overall F-scores, though slightly different from our metric, found a similar pattern of results: the Pursuit model outperforms the Cross-situational model. The difference between models in our results is substantial, with Pursuit correctly learning 57% of the meanings for monosemous words, while Cross-situational only learned 43%. The next set of model results we consider are the results for multi-meaning words, shown in the second and third panels of Figure 4.5. Here, we found that the effect was reversed; Pursuit was only able to learn even a single meaning for 29% of the polysemous words, while the Cross-situational model was able to learn at least one correct meaning for 57% of the polysemous words.

Finally, and most relevant to this chapter's behavioral experiment, we calculated the percent of polysemous words for which more than one meaning was learned (see the last panel of Figure 4.5). We found that the Cross-situational model's performance degraded significantly, as it was only able to learn more than one meaning for one of the seven polysemous words. In comparison, however, the Pursuit model performed even worse, as it was unable to learn even a single secondary meaning. These results imply that hypothesis-testing models such as Pursuit may perform well given less naturalistic input, which does not include multiple meanings. They further suggest that associative models may be comparatively more robust to this phenomenon. However, neither model distinguishes itself as particularly effective at learning multi-meaning input.

Discussion

To summarize, we investigated whether toddlers could recognize multiple meanings for polysemous words, and whether they were able to guess new related meanings upon seeing possible extensions. We tested monolingual English-learning toddlers on real polysemous English words, which happen to generalize differently in an unfamiliar language (Spanish). We found evidence for both (1) experience-driven recognition of polysemous meanings in two-year-old children in the English meanings condition and (2) inference of meanings in the Novel extension condition. That is, results confirmed that children showed above-chance accuracy in identifying the two English meanings of the word and a tertiary meaning that is novel in English (but that exists in Spanish). Critically, performance on English meaning trials (containing English meanings, which our participants had been exposed to) was higher than on Novel extension trials (similar meanings, which our participants would have to infer). This provides evidence that toddlers have already learned multiple meanings for the polysemous words we tested, rather than simply guessing conventional meanings on the basis of a single stored meaning or a vague or underspecified representation.

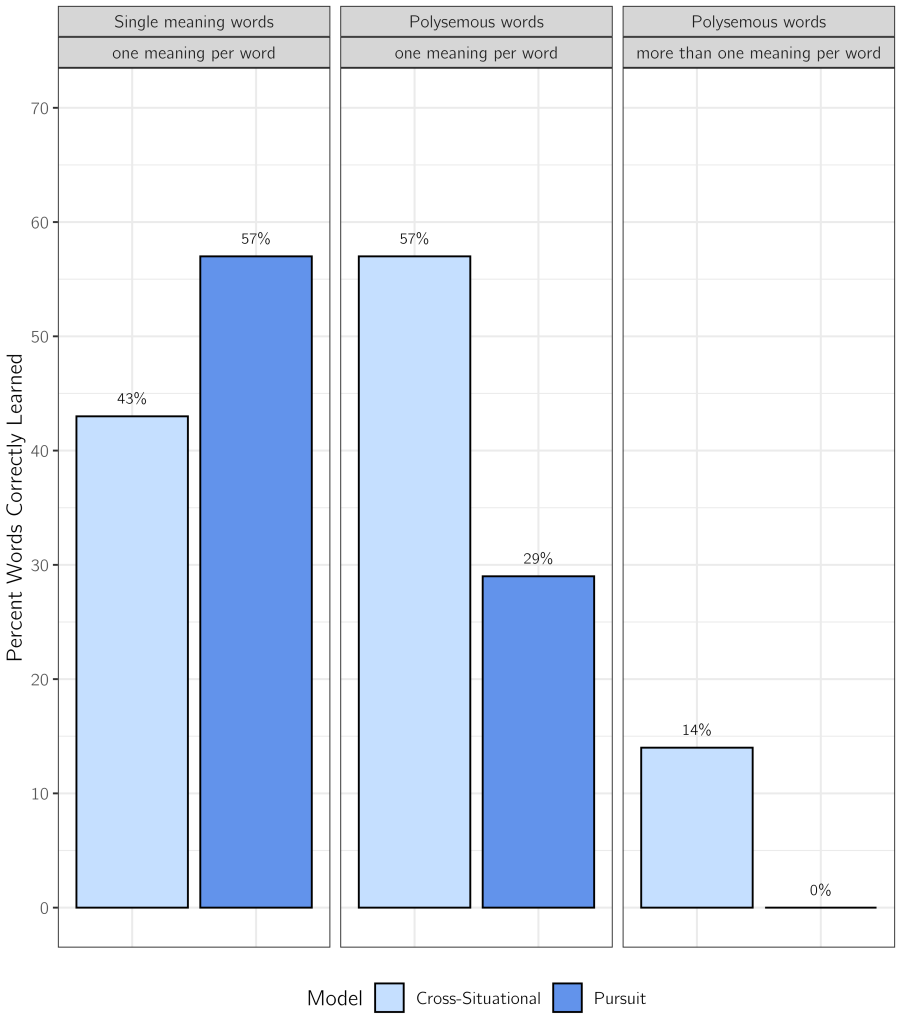
We next tested two predominant theories of early word leaning, Cross-situational word learning and Pursuit. We implemented these models using naturalistic, polysemous language input which reasonably represented what participants in our study had been exposed to. We compared the models' performance on learning polysemy to a baseline performance on single-meaning words. Model comparison revealed that: (1) both models showed degraded performance learning multiple meanings from polysemous word input, and (2) the Cross-situational model, an associative learning model, slightly outperformed Pursuit, a hypothesis-testing model, complicating past work which had tested performance on single-

meaning input. Together, these results suggest that a model which can track multiple associations (as in Cross-situational learning) rather than primarily testing individual hypotheses (Pursuit) may fare better in learning polysemy. Moreover, due to low performance on polysemy by both models, a new approach which can benefit from generalizations across meanings may be needed.

4.3 Chapter summary

These results take an important step towards understanding how the rich and varied meanings of 40-80% of real words are learned. Toddlers' performance on English meaning trials provides evidence that even 2 1/2 year-olds already represent and recognize multiple meanings for a single wordform. At the same time, our findings demonstrate that children are capable of flexibly inferring an additional potential meaning of a familiar word on the basis of the relationship between the new and familiar meanings. We additionally ran model comparison experiments using CHILDES input, finding that a primarily associative model is slightly more robust to polysemous input than a hypothesis-testing model. However, likely due to the competition between meanings which both approaches use, neither approach fared particularly well. In the following chapters, I will further explore cognitive mechanisms which may offer an alternative to purely competition-based models.

Figure 4.5: Model performance on CHILDES input. These results show that learning of single-meaning words, shown in the first panel, is relatively high. The second panel shows the percent of polysemous words for which the models correctly learned at least one meaning. Finally, the third panel shows the percent of polysemous words for which the models successfully learned more than one meaning.



5

Children on the autism spectrum are challenged by complex word meanings

Children on the autism spectrum are recognized to commonly face delays and ongoing challenges in language learning (Henry et al., 2018; Kim et al., 2014). Differences related to joint attention and social skill contribute to these challenges (Kuhl et al., 2005; Mundy

et al., 1990; Paul, 2003), and language therapies therefore tend to focus on improving these abilities (Paul, 2003; Schreibman, 2000). Yet the emphasis on the impact of social skills overlooks the potential role of differential cognitive effects on language development in those on the spectrum (cf. Eigsti & Schuh, 2016; Ozonoff & McEvoy, 1994; Schuh et al., 2016). We are particularly interested in the fact that individuals on the spectrum tend to have difficulty identifying relationships among individual instances to form more general categories. For instance, Plaisted et al. (1998) demonstrated that children on the autism spectrum were less successful than NT children at identifying new instances of a category of dot patterns, and instead tended to treat similar dot patterns as entirely novel*.

Relatedly, children on the autism spectrum show a reduced tendency to sort by gestalt principles (Brosnan et al., 2004), and have been found to generalize categories more easily on the basis of rule-based strategies involving a single dimension (such as shape) than on the basis of multiple dimensions ("red and round") (Klinger & Dawson, 2001; Minshew et al., 1992; Rutherford & McIntosh, 2007). Rather than attending to relationships and generalizations, children on the spectrum tend to devote enhanced attention to perceptual distinctions, resulting in a greater tendency to treat similar instances as entirely new (Molesworth et al., 2005; Mottron et al., 2006; Mottron & Morasse, 2001). This tendency to focus on distinctions over generalizations potentially has broad implications. For instance, learning to predict upcoming events may require us to recognize that a current series of events relates to events we experienced previously. The finding that individuals on the spectrum find it more challenging to make predictions (Sinha et al., 2014) may in part stem from problems detecting similarities and relationships across experiences.

*Portions of this chapter come from Floyd et al., 2021b

In language, the ability to recognize relatedness among individual instances in order to form generalizations is essential (Goldberg, 2019; Tomasello, 2009). Consider the simple case of how individual words are pronounced. Different speakers will pronounce the same word differently depending on their gender, their size, and their dialect, and each person's pronunciation will additionally vary, depending on speech rate and context (Elman, 2009). Although listeners only hear individual instances, we must be ready to recognize a new instance as a member of the same word category in order to recognize it. Similarly, the ability to interpret utterances in which familiar words are used in new ways (e.g., Daddy laughed me off the couch) requires that language learners generalize across utterances and learn the patterns, or grammatical constructions of their languages. One small study compared 7-year-old NT children with youth on the autism spectrum (ages 8-13) on their ability to generalize from examples of a newly learned novel grammatical construction to new instances of the construction (Johnson et al., 2012). While both groups displayed comparable above-chance memory of the original examples, the children on the spectrum lagged far behind in their ability to interpret new instances. Additional support for the idea that challenges in generalization may affect the overall language profile of children on the autism spectrum comes from a tendency to rely more on stored verbatim utterances, and show less productivity in figurative language (Brosnan et al., 2004; Molesworth et al., 2005; Plaisted et al., 1998; Rutherford & McIntosh, 2007). We here explore the idea that challenges in recognizing relationships across instances may affect vocabulary learning in individuals on the autism spectrum.

Learning the meanings of words is critical for successful communication and classroom learning (Hawa & Spanoudis, 2014). An under-appreciated challenge involved stems from

the fact that a single label is frequently associated with multiple meanings (Piantadosi et al., 2012). Homonyms are words that are associated with two or more unrelated meanings, such as bat, which can refer to either a flying mammal or a wooden stick. Vastly more common are polysemous words, which are associated with distinct but related meanings (Britton, 1978; Brocher et al., 2017; Dautriche et al., 2016; Durkin & Manning, 1989b; Fellbaum, 1998; Fillmore & Atkins, 2000; Floyd & Goldberg, 2020; Geeraerts & Gentner, 1993; Lakoff, 1987; Lewandowska-Tomaszczyk, 2012; Navigli & Ponzetto, 2012a; Srinivasan et al., 2019). For example, cap can be used to label a baseball cap, a bottle cap, or a pen cap, each of which involves a distinct shape, material, and purpose, but all of which refer to something that tightly covers something else. While only about 4% of English words are homonymous, the percentage of polysemous words is an order of magnitude greater (40-80%) (Dautriche et al., 2016; Fellbaum, 1998; Rodd et al., 2012).

Recent work has found that neurotypical (NT) children and adults find it relatively easy to learn polysemous word meanings and can often interpret a new related meaning upon a single encounter (Floyd & Goldberg, 2020; Srinivasan et al., 2019). In a direct comparison in which NT participants were exposed to several novel polysemous words or several novel homonyms, both 4½-7 year-olds and adults displayed a marked advantage in learning polysemy over homonymy, and both NT groups demonstrated strong retention of the polysemous word meanings after a week delay (Floyd & Goldberg, 2020). This finding motivates the current work, which asks whether youth on the autism spectrum exhibit an equivalent polysemy over homonymy advantage in word learning.

Evidence that challenges in generalization affect vocabulary learning in children on the autism spectrum comes from research on the "shape bias," which is the well-documented

tendency for word-learners to extend a word that refers to a solid object for use with other solid objects that share the same shape (Landau et al., 1988; Smith et al., 2002). Evident in neurotypical children by the age of 2, the bias to privilege shape over color, texture or size has been found to be absent in 3 and 4 year-olds on the autism spectrum (Tek et al., 2008). At the same time, there is a good deal of individual variation in the use of a shape bias among children on the spectrum (Tovar et al., 2019), and a longitudinal study with more participants has found that highly verbal participants on the spectrum do eventually develop it (Hartley et al., 2019; Potrzeba et al., 2015). We suggest that the delay in learning the shape bias may stem from individuals' greater challenge generalizing across exemplars.

There is evidence that individuals on the autism spectrum make less spontaneous use of polysemous words than neurotypical adults (de Villiers et al., 2013), and can find it challenging to generalize phrases to flexibly and appropriately apply them in new contexts (Brosnan et al., 2004; Mottron & Morasse, 2001). For instance, it may be harder to understand what *boiling mad* means if the metaphorical relationship between heat and anger is not appreciated. To the extent that youth on the spectrum privilege differences between concepts over relationships among them, we predict that the learning advantage for polysemy over homonymy, evident in neurotypicals, will be attenuated in this group. A baseball cap and a bottle cap differ, after all, in shape, material, size, and function, and these differences may interfere with the recognition of their related covering functions.

A narrower focus of attention and concomitant challenges forming generalizations may affect vocabulary learning, given that a plurality of words involve complex constellations of related meanings. Therefore, we hypothesize that youth on the autism spectrum will not benefit from the same polysemy over homonymy advantage in word learning that has been

found in neurotypical word learners.

5.1 Experiment 4

Design, sample sizes, age ranges, and basic analyses were preregistered at AsPredicted.org. Each child was exposed to 4 novel words in a polysemy condition and 4 novel words in a homonymy condition (8 words in total). Each word was assigned 3 distinct meanings, represented by distinct objects, which were either visually related (polysemy) or unrelated (homonymy) to one another (see Chapter 3 for example stimuli).

After brief exposure to the novel words and images of objects, along with a larger set of distractor images, children were tested on how accurately they identified the 3 target meanings for each novel word, from sets that additionally included 5 previously witnessed distractors. Children who were available were recruited again after a week delay and retested on the same task, without re-exposure.

Methods

Participants. We recruited 40 youth on the autism spectrum ages 7-14 (M =11.68 yrs, SD = 1.82; 21 male) from various schools and extracurricular programs. Each had received a clinical or school diagnosis of ASD, and were reported to be "verbal" or "highly verbal" by a parent or guardian. Thirty-six children in this group also took the PPVT-IV, a measure of receptive vocabulary to determine the age equivalencies of their vocabulary. (The order of the main and PPVT tasks were counterbalanced.) 60 neurotypical (NT) children, ages 5-12 were also tested (M = 7.74 yrs, SD = 1.67, 33 male). NT children were recruited through a local summer program. All children were given a book for participation.

Novel Words. The 4 polysemous novel words were constructed such that one "proto-

typical" meaning contained two distinguishing visual features: one feature was shared by a second meaning, and the second feature was shared by a third meaning. That is, no single feature held of all three meanings that could distinguish them from distractors. Stimuli were created in this way in order to be more ecologically valid, since a word's meanings typically do not share necessary or sufficient features (Fillmore & Atkins, 2000; Jackendoff, 1988; Lakoff, 1987).

Exposure. Each participant was exposed to a total of 24 meanings across the polysemy condition (4 words, 3 senses apiece) and homonymy condition (4 words, 3 senses apiece), and 40 distractor images. The 8 novel word labels, e.g., kaisee, veebo, obeyed English phonotactic constraints. Each condition involved 2 exposure videos, containing a randomized stream of 16 novel objects, presented one at a time. Each 2-minute video included 2 novel words and their associated 3 meanings as well as 10 distractor images; three of the six novel meanings were labelled by one word ("This is a kaisee"), and 3 other new novel objects were labelled by a second word. All images were witnessed once, except the "prototypical" object, which was seen twice (exposure matched across conditions). When a distractor was on screen, a brief tone instead of a label was played. The only difference between the homonymy and polysemy conditions was whether the novel words were assigned 3 related meanings (polysemy) or 3 unrelated meanings (homonymy). The order of conditions was counterbalanced across participants, and pairings between labels and novel object sets were also counterbalanced.

Test. On each trial, children saw a display of 8 images including 3 target senses of a word and 5 distractors that had been witnessed during the same exposure video. They were then asked, for example, to "Pick three kaisees" (the order of the 4 novel words within each con-

dition was counterbalanced). Children were tested in a quiet area and wore noise-blocking headphones. The experiment was displayed on a monitor. Children selected answers using a wireless mouse, except if unable to operate it, in which case the child indicated options by pointing.

Results

As preregistered, we entered the average accuracy per subject for each condition into separate paired t-tests within each group (NTs and autism spectrum). As predicted, we found that NT participants performed significantly better in the polysemy condition than the homonymy condition ($t(59) = 5.56$, $p = 6.77e\text{-}07$). Critically, the group on the spectrum showed no difference between conditions ($t(39) = 0.39$, $p = 0.69$). These results were confirmed with secondary preregistered analyses using mixed effect models. To ensure that the polysemy advantage was not solely detected in the NT population due to a difference in sample size, we also subsampled 40 NT participants closest to the 40 ASD participants in homonymy performance using nearest-neighbor matching, and the result was again confirmed: the 40 NT participants show a significant polysemy advantage $t(39) = 5.25$, $p = 5.72e\text{-}06$, while the ASD participants do not $t(39) = 0.39$, $p = 0.69$.

In order to ensure that our analyses were robust, we ran a stronger test on the condition difference between our two groups. Using model comparison (Voeten, 2020) on the maximal converging model with random slopes and intercepts (Barr et al., 2013), we found that there was a significant interaction between group and condition, in which participants on the spectrum showed significantly lower performance in polysemy (vs. homonymy) as compared to their neurotypical peers ($\beta = 0.35$, $t = 3.1$, $p = 0.002$).

Nineteen NT children and 20 youth on the spectrum were available after a week delay

and were retested on the same test trials in the same order, without any additional exposure (see Figure 5.1). We again found that performance in the polysemy condition differed significantly from homonymy for the NT group ($t(18) = 3.22$, $p = 0.005$). Youth on the autism spectrum showed marginally better accuracy in polysemy over homonymy at the second time point ($t(19) = 1.72$, $p = 0.10$). However, the marginal advantage in the ASD group was not reliable when controlling for items and subjects using multilevel modeling ($\beta = 0.16$, $t = 0.929$, $p = 0.39$), while the NT polysemy advantage was robust ($\beta = 0.53$, $t = 3.21$, $p = 0.01$).

The current results suggest that youth on the spectrum face an ongoing challenge: while TD performance in polysemy increased slightly with age ($\beta = 0.11$, $t = 2.83$, $p = 0.01$), we found no evidence of an increase with age in the ASD group using the same preregistered analysis ($\beta = 0.01$, $t = 0.26$, $p = 0.79$). The challenge in learning complex meanings is not likely due to failure to learn words with single meanings: of the 36 ASD participants tested on the vocabulary measure, their vocab-equivalent age was greater than the NT children's chronological age (ASD vocab age equivalent: $M = 10.1$ yrs, $SD = 5.07$ vs. $M = 7.74$, $SD = 1.67$; $t(94) = 3.33$, $p = .001$).

Discussion

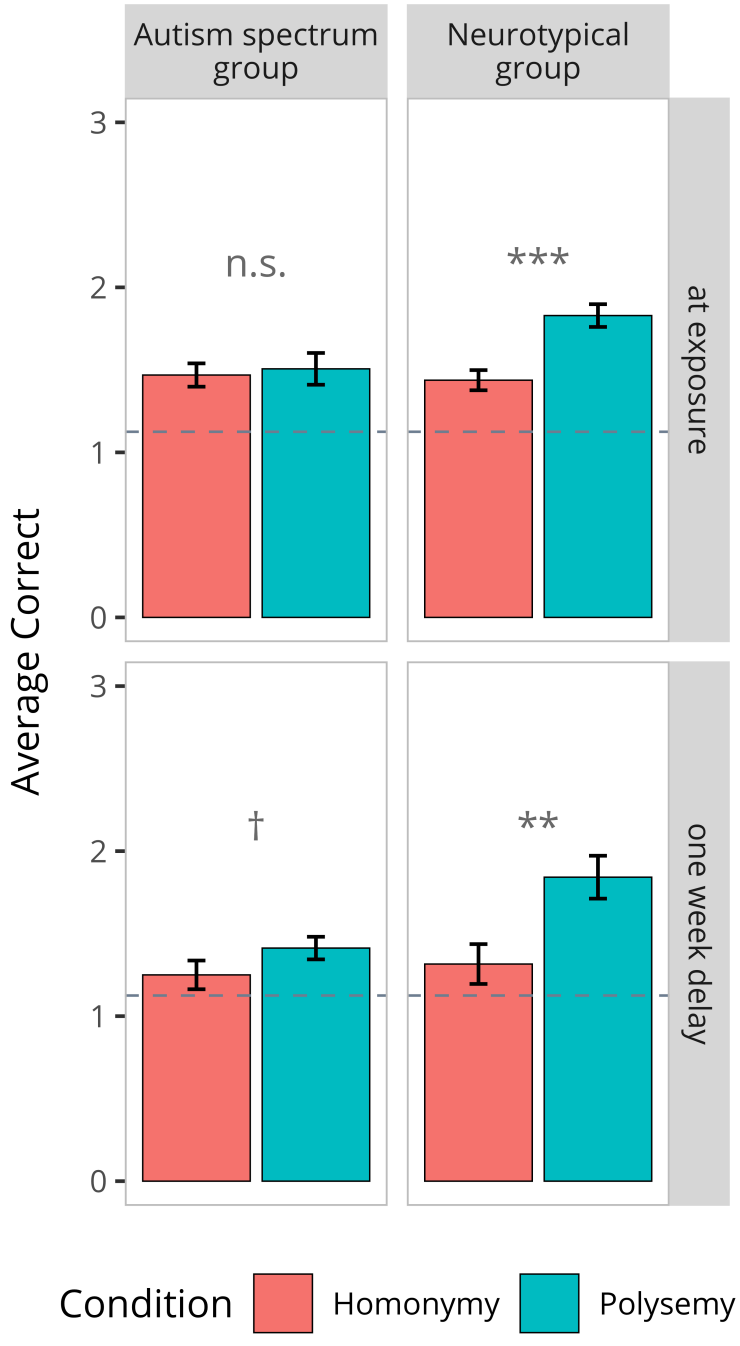
Neurotypical children and adults find it much easier to learn and remember related meanings of a word compared to unrelated meanings (Floyd & Goldberg, 2020). The current results replicate this polysemy over homonymy advantage for NT children in a new within-subjects paradigm through exposure to 4 polysemous and 4 homonymous novel words, with three distinct meanings apiece. Notably, the ASD and NT groups showed equivalent, above-chance performance in the homonymy condition, demonstrating compa-

rable understanding of the task and attention level. But strikingly, the youth on the autism spectrum failed to show the same robust polysemy over homonymy advantage enjoyed by their NT peers. Since a plurality of words are polysemous in natural languages, this finding predicts vocabulary learning to be much harder for individuals on the autism spectrum. While deficits in social skill and joint attention constitute the focus of many interventions aimed at improving communication and language skills in this population (Mottron et al., 2006; Paul, 2003; Rutherford & McIntosh, 2007), the current work emphasizes that learning to understand and use words in contextually appropriate ways may be compounded by challenges in forming generalizations.

5.2 Chapter summary

Neurotypical children find it relatively easy to learn that glasses can mean both "drinking cups made of glass" and "eye-glasses" because of the relationship between the two meanings. But children on the autism spectrum are not advantaged in the same way by the relationships between distinct meanings; they may find it no easier than learning that glasses mean "eye-glasses" and "porcelain mugs." A recognition of the challenge that polysemous words present to youth on the autism spectrum may be useful for interventions aimed at improving communication skills in this group. In particular, support may be needed for youth on the autism spectrum to recognize that a word witnessed in a particular context with a particular meaning can also be used in a different context with a related but distinct meaning

Figure 5.1: Top row, NT children and children on the spectrum showed equivalent, above-chance accuracy on novel homonyms (pink). While NT children were significantly more accurate on polysemes, children on the spectrum displayed no difference (blue). Bottom row, performance after 1-week delay shows same pattern of results.



6

Modeling rules and similarity in colexification

English speakers can often use the name of a species to refer to the animal's meat (e.g., she fed the *fish*/she ate *fish*). But speakers also face uncertainty: the meat of a *cow* is referred to as *steak* (or *beef*). If offered the meat of a zebra at a restaurant, it is unclear whether an En-

glish speaker should apply a label for a similar concept, referring to it as *steak*, or follow an animal-for-meat generalization or "rule', labeling the meat *zebra*. Either choice leads to an extended use of a familiar word (*steak* or *zebra*) . In fact, a a plurality of familiar concepts are **colexified**, sharing a word label with at least one other concept (Zipf, 1945b). Like the multiple meanings of *fish* and *shower*, most colexified words name related or **polysemous** meanings (Bréal, 1897; Durkin & Manning, 1989a; Navigli & Ponzetto, 2012b; Rodd et al., 2002a; Schmitt & McCarthy, 1997), rather than unrelated or **homonymous** meanings(Dautriche, 2015a). In this chapter, I ask what underlies the relationships between multiple meanings in the minds of language users*.

Polysemous meanings are by definition *related* meanings, so they do not generally lend themselves to a single definition of necessary and sufficient attributes. Instead, they require learners to somehow form a complex category of relationships between labels and meanings (Wittgenstein, 1953). A quick examination of cross-linguistic differences confirms that there are a multitude of cases of polysemy that are not universal(Fillmore & Langėndoen, 1971; François, 2008; Rzymski et al., 2020). For example, English's *straight*, French's *droit*, and Persian's *rost* can all be used to mean both "rectilinear" and "honest" (e.g., *straight talk*). The semantic extension from "rectilinear" to "honest" is motivated by a general metaphorical relationship between speaking directly and speaking honestly, but not all languages use the same word to mean both "straight" and "honest" (e.g. Spanish does not), so this is something that English, French, and Persian speakers must learn. English additionally uses *straight* to mean, "undiluted" and "heterosexual", while neither French nor Spanish do (François, 2008). Differences across languages has been used as evidence

*Portions of this chapter come from Floyd et al., 2021a

that people must in fact learn the extensions of their language as a set of conventions based on language input (Lehrer, 1990b; Murphy, 2004a). It remains unclear, however, just how idiosyncratic polysemy is across languages.

Past research has looked for recurring patterns of meaning extension or "rules" which might be used to predict colexifications, at least probabilistically. A rule would allow a new meaning to be learned and represented by applying the rule to an existing meaning (Fauconnier, 1994; Nunberg, 1979b). Rule-based extensions are predicted to reappear across languages (Copestake & Briscoe, 1995; Lakoff & Johnson, 2008; Ostler & Atkins, 1991; Pustejovsky, 1998). Cross-linguistic work tested 14 languages for instances of 27 English patterns of polysemy (such as "animal-for-meat") (Srinivasan & Rabagliati, 2015). English speakers who were fluent in one of the other 14 languages were presented with with a single example of a rule in English and asked to come up with further examples in their other language. In every instance, there was at least some vocabulary which followed the English extension rules tested. However, there was also considerable variation regarding which particular vocabulary items the rule could apply to. For example, while English can be said to have a "material-for-artifact" rule (e.g. made of glass, a glass of milk; made of tin, a tin of fish), other languages apply this extention to materials other than those found in English.

Other evidence suggests a role semantic relationships that were less general in that they are only relevant to one or two concepts in a language (hereafter, "similarity") in colexification (Wittgenstein, 1953). One study predicted meaning change across time in English, (Ramiro et al., 2018b), using a nearest-neighbor chaining algorithm over Word2Vec embeddings in order to predict the emergence of new, similarity-based meaning extensions for an existing form at later timepoints. This work reported that similarity played a strong

role in predicting which sense a word acquired. Yet this work did not compare results with a rule-based account. Also, by predicting historical changes in English rather than directly testing human performance on novel colexifications, potential confounds may arise. In particular, historical changes may face pressures from many other sources such as environment, colonialism or other language contact (Gordon, 2004; Kirby et al., 2016; Thompson et al., 2020; Wiseman, 2015), making it difficult to draw conclusions about the role of similarity in the minds of language users.

One study that compared rule-based and similarity-based extensions found that English speakers' acceptance of artificial, English-based colexifications was higher in a set of rule-following cases, than cases judged to be similar (Rabagliati et al., 2010). However, the study relied on test cases based in English patterns, and did not test the likely possibility that the two mechanisms (rules and similarity) could work in tandem. As Ramiro et al. and Rabagliati et al. both investigated how English colexifications or English colexification patterns, the question of how learners contend with novel colexifications remains open.

Especially relevant to the present topic is a recent finding that conceptual relatedness, operationalized via word association judgments and similarity measures on Word2Vec vectors, is a strong predictor of colexification across languages (Xu et al., 2020). This work focused on predicting actual colexifications across languages, and established a conceptual basis for attested, high-frequency colexifications which they attribute to cognitive economy—more closely associated concepts are more likely to share a label in languages. However, this work did not investigate the contribution of rule-based polysemy. And though it is surely the case that speakers of language must learn the polysemous conventions of their language, this study cannot inform our understanding of the word learning process. Further, the

colexifications in existing languages are often subject to historical events and language contact, complicating our ability to use them to understand word learning. Therefore, the present work investigates a distinct, but related question: When presented with novel word meanings, to what extent do learners rely on productive, rule-based extensions vs. general semantic similarity?

Because language is rife with colexifications between concepts, the present work aims to determine the role of two candidate mechanisms in the human mind. We test whether rule-based generalizations predict how learners expect concepts to colexify, e.g. a rule would predict that learners would expect the concepts "foot" and "leg" are likely to colexify because they follow a part-for-whole pattern. We also investigate the contrasting or potentially complementary proposal: that similarity predicts expectations about colexification (e.g. the concepts "float" and "swim" are likely to colexify because their semantic similarity is high). We suggest a third possibility: that humans simultaneously use *both* rules and similarity in structuring word meanings. Past work has used Bayesian modeling to compare rule and similarity-based learning in the domain of numerical concepts (Tenenbaum, 2000). The Bayesian modeling approach is well-suited to investigating these questions, as it offers a mechanism-agnostic approach to formalizing multiple kinds of hypotheses, such as numerical and semantic similarity, as well as rule or pattern-based structures. In this vein, we propose and evaluate three Bayesian models of colexification: one based in rules gathered from extant literature, one based in similarity as measured through word embeddings, and a hybrid rule-and-similarity model which generates predictions from a mixture of the two models. Based on past literature on concept learning, we predict that the hybrid model will correlate best with human expectations for colexifications in natural language.

6.1 Experiment 5

To test this prediction, we asked adult, English-speaking participants to rate the probability of pairs of concepts being expressed by the same label in another language. The concept pairs were selected to be commonly attested colexifications in a database of over 3K languages, but novel to our English-speaking participants. We then correlated the human ratings with the three models' predictions. Critically, the participants' and models' ratings depend on the prior distribution of both rule-based and similarity-based colexifications in their experience. Therefore, the task was preceded by a warm-up phase, in which participants were exposed to a set of frequent colexifications from outside their language (randomly sampled without replacement from the full set used in the main task), so that the participants' prior expectations could be aligned with the models' priors.

Table 6.1: Examples of hypotheses from English as well as our task (non-English colexifications).

Hypothesis type	Example
Animal-for-meat (rule)	***feed a chicken, eat chicken*** [English]
Metaphor (rule)	***a beautiful face, face my problems head-on*** [English] ***voice, perspective*** [English]
Cause and Effect (rule)	***rent to, rent from*** [English] ***move quickly, fast*** [other languages, task item] ***drop, fall*** [other languages, task item]
Part-for-whole (rule)	***walk down Wall Street, Wall Street panicked*** [English] ***hand, arm*** [other languages, task item]
Similarity	***old (aged), old (expired)*** [English] ***float, swim*** [other languages, task item] ***grandson, nephew*** [other languages, task item]

Methods

Participants Adults (n = 60) were recruited from Amazon's Mechanical Turk via CloudResearch (Litman et al., 2017a) and compensated $1.20 to perform 100 ratings, preceded by the priors warm-up phase which exposed them to 70 other colexifications. 100% of participants who completed the task were approved for payment and no data were excluded.

Stimuli We obtained natural language colexifications from the Database of Cross-Linguistic

Colexifications (CLICS), an inventory of 2,919 concepts from 3,156 languages across 20 different language families (Rzymski et al., 2020). Because our intent was to test how English speakers structured novel colexifications, we extracted the 300 most frequently colexified meanings across languages, excluding any which also occurred in English using the CLICS colexification set for English. However, the English colexification set did not span the full list of concepts, so a coder reviewed and removed an additional 55 colexifications from the top 300 which were identified as attested in English, leaving 245 pairs of novel colexified concept pairs. From this set, we randomly selected 70 colexifications to use as exposure items in the warm-up phase which was run to align the participants' experience to the model's training. The remaining 175 colexifications were used for the main task. In order to estimate how participants rated both attested and non-attested colexifications, an additional 175 pairs were created by scrambling concepts named in the 175 attested colexifications.

In order to determine which type of hypothesis the colexified pairs corresponded to, two coders were trained to recognize rule-based extensions based on past literature (Lakoff & Johnson, 2008; Smith et al., 2002) and instructed to code each pair of colexified concepts used in the experiment. Whenever a pair did not correspond to one of the 32 rules, it was categorized as a similarity-based colexification.

Procedure The study consisted of two phases. First, in the priors exposure phase, participants were shown a randomly-selected set of 70 colexifications in order to align their experience in the task with the basis for the models' priors. Participants were shown two concepts from a colexification (such as "leg" and "foot") and told that a foreign language uses the same word to express both concepts. On the next page, they were prompted with

one of the concepts from the colexification they were just exposed to ("foot") and asked to select the corresponding concept ("leg") from among two distractors; they could not proceed to the next exposure until they correctly identified the target concept.

In the main phase of the task, participants were exposed to a randomly-selected 100 pairs of concepts, 1 pair per page, and prompted to rate the probability each pair of concepts being colexified. That is, participants were asked, e.g. "How likely do you think it is that a foreign language would use the same word to express the concept LEG(noun) and the concept FOOT(noun)?" Participants then selected a value from 0 to 100 on a slider, and this rating was compared with model predictions for each item.

6.2 Experiment 5 modeling study

We compare three types of Bayesian models. A rule-based model uses productive patterns of colexification identified in past research (see Experiment). A similarity model used two methods for measuring similarity: cosine similarity between the two colexified concepts' word embeddings from Sentence-BERT, a state-of-the-art transformer model of language (Reimers & Gurevych, 2019), and cosine similarity between the concepts' Word2Vec embeddings (Mikolov et al., 2013), which relies on a much simpler algorithm and less training. This was done in order to determine that the similarity models' performance was consistent, independent of the quality of the word embeddings. Finally, we present a hybrid model which represents a mixture of both rule and similarity models, again separately testing this model using both measures of similarity. We used each model to generate a prediction for the probability that two meanings would be colexified, and this was compared to human predictions for the same pair of meanings from the Experiment.

Rule-only model Each colexification used in the experiment was coded for which hypothesis type (rule or similarity) it corresponded to, and in the case of a rule-type colexification, which particular rule hypothesis (see Table 1). Model predictions (the product of the prior and likelihood for each data point) were then compared with z-scored human ratings from the Experiment.

Formally, the hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$, consisted of N rules, each represented as the set of valid colexifications under the rule (determined by trained coders; see Experiment). Rules were constructed from $N = 32$ extensions believed to be productive and general across languages in past work (see Table 1 for examples).

The prior probability $p(h_i)$ was defined as the proportion of colexifications that adhered to each rule h_i, using the 70 items from the Experiment's priors based on the warm-up exposure.

The likelihood function was binary, based on whether or not the example followed a rule and given by:

$$P(C|h_i) = \begin{cases} 1 & \text{if } C \text{ is consistent with } h_i \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

where a colexification C is an unordered pair $C = C_1, C_2$ of meanings.

Given this prior and likelihood, the probability that a colexification would occur under any rule is given by:

$$p(C|R) = \sum_{i=1}^{n} p(C|h_i)p(h_i). \tag{6.2}$$

In the rule-only model, $p(C|R) = 1$ (as the hypothesis space consists of rules only), but it captures the degree that a colexification is supported by rules in the hybrid model.

Similarity-only model

The similarity-based model consists of a single hypothesis that expresses the probability that two concepts would be colexified based on their semantic similarity, and its prior was therefore set to 1. The likelihood of a colexification under the similarity rule is given by:

$$P(C|S) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{Sim}(C_1, C_2))}}. \tag{6.3}$$

This likelihood consists of a softmaxed linear model over similarity between concepts. The Sim function was computed via cosine similarity between the embeddings of C_1 and C_2, obtained using Sentence-BERT pretrained embeddings (Reimers & Gurevych, 2019) in one version of the model and Word2Vec pretrained embeddings in another (Mikolov et al., 2013). The final softmaxed linear model was obtained by fitting a logistic regression fit to the full set of colexifications used in the task via Ordinary Least Squares.

Hybrid rule-and-similarity model

Our main model combines predictions from both similarity and rule models described above, using a mixture given by:

$$p(C) = p(C|R)p(R) + P(C|S)p(S) \tag{6.4}$$

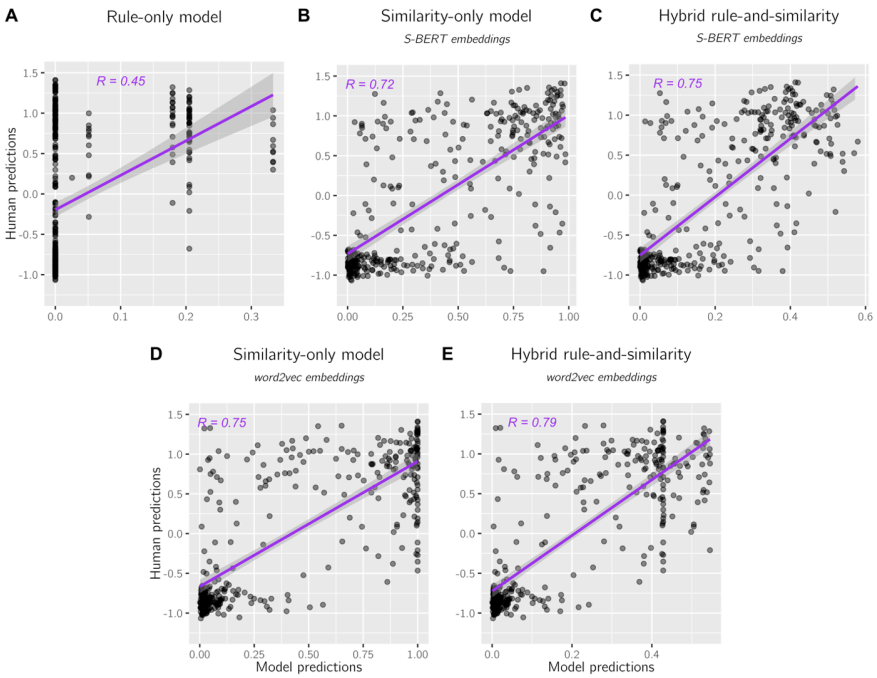
where $p(R)$ and $p(S)$ were scaled so that $p(R) + p(S) = 1$ using the distribution from the stimuli presented in the prior exposure phase (see Experiment), leading to $p(R) = 0.57$ and $p(S) = 0.43$.

Results We correlated model predictions for each pair of concepts with z-scored human predictions for that item. Though all models showed some evidence of predicting human behavior (see Table 2), the hybrid rule-and-similarity model consistently and significantly outperformed the next-best performing similarity-only model in both versions ($z = 3.19$, $p = 0.001$ with S-BERT, $z = 2.68$, $p = 0.01$ with Word2Vec). Our prediction, that the hybrid rule-and-similarity model would best correlate with human judgments, was borne out in the data. However, panels C and E of Figure 6.1 show quite a few items which are outside of the hybrid model's predictions (panels C and E). For example, the colexification between "dye" and "paint", which is attested in many languages, was given quite high ratings from the participants, but less high predictions from the similarity model. And, even though the two concepts follow a productive rule ("same function"), the frequency of this rule in the prior estimation data set was quite low, meaning that the combinations of low similarity and low rule-based hierarchical prior generated a much lower prediction from the model than the probability assigned by participants.

We also found that the similarity-based model correlated fairly well with human judgments (panels B and D). As in the hybrid model, there is still a cluster of attested colexification items which neither the model or humans think are likely, suggesting that even though similarity offers a gradient prediction, the similarity likelihood function assigns very low probability to less similar items.

In the top right-hand corner of panels B and D, predictions from the similarity-based model shows that there is a rather large cluster of colexifications with high similarity ratings. In comparison to the hybrid model (panels C and E), this suggests that the hybrid model may have benefitted from the hyperprior scaling down the similarity-only model's

Figure 6.1: A: rule-only model weakly predicts variation in human responses. B: similarity-only model using Sentence-BERT embeddings performs noticeably better. C: a mixture of rules and similarity using Sentence-BERT embeddings correlates even better than similarity alone with the same embeddings. D & E show the same similarity-only and hybrid models, instead using Word2Vec embeddings, which had the same pattern of results (hybrid model > similarity-only).



high predictions, introducing more variation and increasing the correlation with human predictions.

The rule model showed the weakest correlation with human performance (panel A). One possible reason for this is that, because each rule represents a categorical relationship, a model based in rules is ill-suited to capture the variation in participants' finer-grained predictions for each item. A second explanation we considered was that participants may not have been exposed to examples of each of the rules during the priors exposure warm-up, since they were only exposed to 70 colexifications. However, every rule type which appeared in the main task did appear in the prior estimation data, so participants would have

had experience with colexifications from each rule. Nonetheless, it is possible that the frequency of these examples was not high enough in the prior phase to sufficiently reflect the frequency of that rule type in the main task.

Table 6.2: Pearson's correlations between human and model predictions

Model type	Correlation
Rule-only model	0.45
Similarity-only model (w/Sentence-BERT embeddings)	0.72
Similarity-only model (w/Word2Vec embeddings)	0.75
Hybrid rule-and-similarity model (w/Sentence-BERT embeddings)	0.75
Hybrid rule-and-similarity model (w/Word2Vec embeddings)	0.79

As described in the Models section, we ran both the hybrid and similarity-only models using Word2Vec embeddings as well as Sentence-BERT. This was done to test the possibility that the similarity model's results were not dependent on the state-of-the-art transformer architecture or extensive training used to create the S-BERT embeddings. However, we found the same pattern of results with Word2Vec as with S-BERT (see Table 2). In fact, the correlations between humans and the hybrid rule-and-similarity model as well as the similarity-only model were higher for Word2Vec, suggesting that the superior performance of models which use similarity are not entirely dependent on the sophistication of the similarity measure. It is, however, not intuitive that embeddings from a simpler model with less training such as Word2Vec should deliver superior results. One possible explanation is that these pre-trained vectors were used to create embeddings for the short concepts described

in the CLICS database, which often did not exceed a single word in length. And while Sentence-BERT was trained to build representations for entire sentences, Word2Vec was trained to build representations for words, which are overwhelmingly the semantic unit in our task, rather than phrases or sentences.

Discussion The present study investigated mechanisms which underlie the human ability to share labels across meanings. Specifically, we formalized a label-generalization task as Bayesian inference over hypotheses based in semantic similarity, hypotheses based in rules, and a mixture of the two. We collected behavioral data from 60 English-speaking participants, asking them to rate the probability of novel colexifications based in attested language data. We found that the hybrid model, which combined predictions from both rule-based and similarity-based hypotheses, outperformed each lone model. Additionally, though the hybrid model's correlation with human predictions was significantly higher than the similarity-only model, the similarity-only model's correlation was almost as high, and the rule-based model correlated worst with human judgments.

To represent each concept in the similarity model, we used word embeddings, which are models trained on large text corpora, and then calculated distance between the colexified concepts in this semantic space. The ability of these corpus-based semantic representations to capture human predictions might suggest that at least one of the possible structures underlying colexifications can be learned. Of course, the sheer amount of training required in these embeddings far exceeds what individual learners experience. Critically, however, we tested our models with two different kinds of word embeddings, and found that our similarity and hybrid models correlate just as well when trained on significantly less data with a much simpler algorithm.

A possible explanation for the relatively poor performance of the rule-only model is that our hypothesized rules could not adequately characterize the colexifications tested, especially because the patterns we included were deliberately selected for their novelty to English speakers. However, past research has found that these rules are attested, in some form, in each of the 14 languages they tested (Srinivasan & Rabagliati, 2015). Secondly, we expanded our rule set by adding 5 even more broad and inclusive rule types from additional literature, including "metaphor", "cause-and-effect" and "same function". Moreover, participants were exposed to at least one instance of each rule type that appeared in the main task, during the priors exposure phase. Therefore, we do not think the best explanation for the rule model's low performance is that our candidate rules were insufficient. However, this possibility does open new avenues of research: future work should examine possible systematicity in colexifications outside of English using human coders and/or machine learning techniques.

6.3 Chapter summary

Together, we show the first quantitative modeling evidence that suggests humans structure colexifications using both semantic similarity *and* rules (such as the part-for-whole rule and the semantic similarity between the concepts "leg" and "foot"). This is largely consistent with similar findings from concept learning, which show that human predictions can reflect a mixture of both (Tenenbaum, 2000), though our results suggests a large part of the hybrid model's success is due to similarity. We build on earlier work in semantics by testing more than one simultaneous account, going beyond cases already known to participants, and generating quantifiably-testable predictions which we compare with human data. Our

results shed light on previous findings which have been interpreted as evidence for rules (Srinivasan & Rabagliati, 2015), evidence for rules over similarity (Rabagliati et al., 2010), and as evidence for similarity-based chaining over other category structures (Ramiro et al., 2018b), and are convergent with a recent finding on the role of conceptual relatedness in attested colexifications across languages (Xu et al., 2020). By using a modeling framework that can simultaneously represent multiple mechanisms, we test whether a combination of the two predicts human behavior better than either mechanism alone.

7

Conclusion: The network perspective

Words are fundamental building blocks of language, communication, and culture, and we often naively think of them as straightforward expressions of meaning. However, rather than referring to a single concept, up to 80% of words are associated with multiple, related meanings (**polysemy**; Bréal 1897; Dautriche 2015a; Durkin & Manning 1989a; Fillmore

1992b; Floyd & Goldberg 2020; Rodd et al. 2002a; Srinivasan et al. 2019; Zipf 1945b). The ubiquity of such complex meanings challenges the common intuition that words can be defined or easily formalized, and raises questions about what kind of mechanisms are involved in human language learning.

7.1 Chapter summaries

How early are multiple meanings learned? How do learners track more than one meaning? What kinds of mechanisms of generalization might be involved? In this dissertation, I have approached these kinds of questions using behavioral, developmental, clinical, and computational approaches. In Chapter 1, I reviewed literature on word learning and cognitive development, and argued that current theories of word learning need to be extended in order to account for the naturalistic demands of multi-meaning learning. In Chapter 2, I reviewed relevant literature from cognitive psychology to serve as the theoretical background for a new theoretical approach. In Chapter 3, I found that adults and young children can successfully track multiple meanings of new words, and found evidence to suggest that this learning is long-lasting and not explicable simply by means of *a priori* visual similarity, or generalization of a single feature. In Chapter 4, I confirmed using eye-tracking methodology, that children as young as two years old are already familiar with multiple, distinct meanings of certain words; results also revealed that they are capable of inferring additional meanings in certain contexts on the fly. I then quantitatively evaluated the two dominant theories of word learning on their ability to learn multiple meanings from naturalistic input, and observed only weak learning of multiple meanings in either model. In Chapter 5, I investigated the role of generalization in word learning by comparing youth on the Autism

spectrum with neurotypical children who performed comparably when tasked with learning unrelated meanings of novel words. I found that youth on the Autism spectrum do not benefit from relatedness among meanings of words the way neurotypical children do, likely due to a tendency to focus on specifics rather than generalizations or relatedness among entities. In Chapter 6, I investigated which mechanisms underlie the human ability to assign distinct meanings to the same label; to do this I developed and compared Bayesian models of semantic similarity vs. previously-posited extension rules. In the following paragraphs I provide a more in-depth summary of each empirical study (Chapters 3, 4, 5, and 6).

Summary of Chapter 3. Experiments 1 and 2 tested children and adults ability to rapidly learn and retain multiple, new meanings for novel words. I found that even when a word's meanings relate in complex, idiosyncratic ways, adults and children as young as four years of age are able to learn them from minimal exposure, with retention remaining strong after a week. More specifically, I exposed each participant to four novel words with three novel meanings apiece in two conditions. In the polysemy condition, meanings of each word were semantically related and in the homonymy condition, the meanings of each word were unrelated. Participants took part in both conditions to control for the memorability of meanings. We found that both children and adults spontaneously benefit from relationships across meanings, even when these word meanings relate in multidimensional, idiosyncratic ways, as cognitive linguists have long presumed on the basis of natural language. Importantly, even the young learners showed evidence of long-term retention of polysemous meanings across 2 different timepoints and 2 different paradigms, without re-exposure.

I ruled out several alternative explanations through extensive norming. In particular, the

advantage of polysemy over homonymy was not attributable to *a priori* visual similarity, or to generalization based on a single shared feature. This work offers a new possible explanation, based in language learnability, for the prevalence of multi-meaning words across languages. This phenomenon has typically been explained through the production-based advantage of reusing words that are easy to access and produce (Harmon & Kapatsinski, 2017; Piantadosi et al., 2012). However, this explanation does not predict that most words should have related, rather than unrelated (homonymous) meanings. The data presented in Chapter 2 suggests that polysemous words may be highly frequent not just because they are easy to produce, but because multiple, related meanings are easier to learn, generalize across, and retain.

Summary of Chapter 4. In this Chapter, I used eye-tracking to explore what very young word learners know about words with multiple meanings, and complementary computational modeling methods to investigate how this knowledge is learned and structured. In an eye-tracking study with monolingual English-learning 2-year-olds, I found that participants successfully recognized multiple meanings for nouns (e.g., the meanings 'baseball cap' and 'bottle cap' for the label *cap*). Next, I investigated whether this word knowledge was specific to previously learned meanings or reflected one underspecified representation that could be applied to novel meanings. To do so, I tested participants with novel extensions for the same English words (e.g., the meaning 'lid' for the word *cap*, a natural polysemous extension in Spanish). I found that toddlers' recognition of novel extensions was above chance, but performance on the English meanings was significantly higher. This suggests that toddlers' knowledge of English polysemy mostly reflected learned, distinct meanings. Finally, I examined two dominant theories of word learning – cross-situational word

learning and a hypothesis-testing model – to see if they could contend with multiple meanings. For both models I observed weak learning of multiple meanings. I conclude that word learning theories should be further developed to account for natural complexities between words and meanings.

Summary of Chapter 5. Chapter 3 established that children and adults find it relatively easier to learn polysemous words, as opposed to homonymous ones. In Chapter 4, I further showed that even toddlers are willing to infer an additional new meaning for polysemous words. Therefore, in Chapter 5, I test the role of generalization ability in polysemous word learning by recruiting children on the Autism spectrum, a population known to devote enhanced attention to perceptual distinctions, resulting in a greater tendency to treat similar instances as entirely new. This study tested children on the spectrum as well as neurotypical children on a word learning task that included 8 novel words (4 homonymous and 4 polysemous). Each word had three meanings which were unrelated (homonymy) or related to one another (polysemy). Interspersed during exposure were a total of 40 unlabeled foil meanings. At test, children were prompted to select the target meanings for each word from among a set of foils. Results confirmed that, while the polysemy over homonymy learning advantage replicated for the typically developing children, there was no indication of any such advantage for autistic children. The data presented in this chapter suggests that a generalization mechanism allows typical word learners to benefit from relationships across meanings. Further, it offers a possible explanation for language and vocabulary delays in individuals on the Autism spectrum.

Summary of Chapter 6. Next, I further explored the nature of the generalizations between polysemous word meanings. Using Bayesian modeling, I investigated the relative

contributions of two possible mechanisms—rules and similarity— in predicting humans' expectations for polysemy. I proposed three computational models based on rules, semantic similarity, and a hybrid of the two, and correlated model predictions to human behavior—in a novel task, monolingual English speakers generalized labels across pairs of colexified meanings from other languages. I found that a hybrid model correlated best with human behavior, showing how a framework which combines rule and similarity-based semantic generalizations can outperform either mechanism alone.

In what remains of the final chapter, I will describe more in-depth the components and possible implementations of a novel framework for the study of word learning.

7.2 The network perspective

I propose that understanding the ubiquity of polysemy and the early age at which children learn polysemous meanings requires a new perspective on word learning. Specifically, I suggest that very young learners (and adults) encode a *network* of related meanings during word learning. This network perspective makes three central claims based in data presented in this dissertation as well as extant literature on semantics, learning, and categorization. First, the network perspective is *distributed*– representations of meaning must be distributed rather than atomic, allowing for representation of attributes of concepts and their contexts which may be in common or diverge across multiple meanings. Second, the network perspective proposes that *generalizations* form freely between meanings, based on overlap in representations, *ad hoc* similarity or relatedness, or, later in learning, productive patterns observed between meanings. Finally, the network perspective proposes that word learning is *additive*–when a new meaning is encountered, learners are open to associating

it with an existing label, even if the label is already associated with other meanings, particularly when the new and extended meanings are related to one another. In combination with generalizations between these distributed representations, additive learning enables meanings which share stronger relationships to existing candidate meanings to become associated more strongly with the word form. These three pieces–distributed representations, generalization, and additive learning– work together to create networks of related meanings and their contexts, as words and their meanings are learned throughout the lifespan. Below, I describe each of these three components (distributed representations, additive learning, and generalization) in greater detail.

Distributed representations in the network perspective.

Models of human word learning, reviewed in Chapters 1 and 4 have almost uniformly represented meanings as atomic, with no identifiable attributes or internal structure. This presents a challenge for learning multiple meanings, as even toddlers readily extend known labels to meanings with shared attributes (see Chapter 4). We propose that, rather than encoding or representing atomic word meanings, word learning requires representations with internal structure, in order to capture relatedness or similarity between a word's meanings. For example, a representation such as "baseball cap" includes multiple attributes such as "covering" and "tightly fitting", which are shared by bottle caps, rendering the two meanings of *cap* related. Based on data presented in Chapter 3, we argue that meaning representations themselves are distributed, as many of our novel word meanings extend on a limited number of features. This suggests that learners are guided by similarities between components of meaning, rather than simply tracking a meaning as a whole.

Outside the field of vocabulary learning, distributed representations are common. For

example, feature-based representations are widely used in models of word form disambiguation (Hinton & Shallice, 1991; Joordens & Besner, 1994; Kawamoto, 1993; Kawamoto et al., 1994; Rodd, 2004); this work has tended to emphasize phonological or orthographic features rather than meaning-based attributes. In rare cases, word learning models have included hand-coded semantic features (Fazly et al., 2010), though this approach is difficult to scale and is subject to coder biases in what features are likely to be observed by learners. To overcome these kinds of limitations, work in computational linguistics and natural language processing has relied on distributed meaning vectors by using corpora to learn word embeddings – which can be as as simple as co-occurrence models, to more complex models such as Word2Vec, or even more complex architectures and training sets such as Sentence-BERT (Devlin et al., 2019; Mikolov et al., 2013). Though many of these word embeddings comprise a single point, this point is situated within a high-dimensional space, each dimension of which can be thought of as contributing a distributed component of meaning to the embedding. These approaches to representing meaning offer at least two possible bases for how distributed meanings can be represented in the network approach.

Generalization in the network perspective.

While children's and adults' generalization abilities have long been documented in psychological literature, the phenomenon of generalizing word forms across multiple concepts has received comparatively less attention. I proposed that witnessing a word in a new context is likely to direct learners' attention to features that were associated with the word's meaning in previous contexts. If some of those features are shared by a possible new meaning, the label would then provide an attentional bias toward the new, related referent (as seen in the experimental task of Chapter 4). As each exemplar of a meaning is to some

extent unique, the same process may even be operative at a lower level, for instance, even when we recognize a new dog as a *dog*, or a new door as a *door*. A generalization mechanism over distributed representations allows for distinct but related meanings reinforce one another, or reduce decay in memory, as seen in Chapters 3 and 5.

Past accounts for how two or more concepts share a label have primarily focused on generalizations rooted in productive rules or patterns, as reviewed in both Chapters 1 and 6. The rule-based approach appears to reduce memory demands, as it avoids requiring that multiple distinct, individual meanings are encoded for each form. Without storage of conventional usage, it is worth noting that a rule-only account could endlessly generate incorrect forms, such as "roast cow" or "soda metal" rather than the more conventional "roast beef" or "soda can". Even more importantly, a rule account leaves an enormous amount for abstraction to represent and compute with no clear mechanism for how these abstractions would be acquired. And in communication, simply applying abstract rules such as "has the same shape as an existing meaning" or "has the same material as an existing meaning" could create confusion, as it is not clear which rule to apply when. On the other hand, the kind generalization process outlined in the network approach, in which relationships are formed in response to how words are used, allows for the kind of vast variation in the specific application of polysemy across languages (François, 2008; Heine & Reh, 1984b; Malt, 2010; Murphy, 2004b; Sweetser, 1990; Traugott, 2014). This variation is what suggests that word learning often involves forming generalizations *ad hoc*, once given a common label across meanings. The kind of generalization which is responsive to usage, as outlined above, allows generalizations to form flexibly, informed by conventions and based on a variety of commonalities or relationships.

Additive learning in the network perspective.

It is perhaps uncontroversial that additional, new meanings become conventional through exposure (Copestake & Briscoe, 1995; Goldberg, 2019; Lakoff, 1987; Tuggy, 1993), and the empirical work reported in this dissertation extends this idea to provide evidence that word learning involves incorporating multiple meanings. In the network perspective described here, I argue that the process of aggregating many meanings is fundamental to word learning broadly: as new meanings and contexts arise in the world (such as, for example, the advent of personal computers), existing words are often extended to label these meanings (such as the computer *file*).

The network approach to learning does not require learners to delete, dampen, or suppress meanings of a word when a new meaning is encountered, as is done in existing theories. While competition effects have been shown in online disambiguation, in order to select a unique word or meaning from among alternatives (Hinton & Shallice, 1991; Jo-ordens & Besner, 1994; Kawamoto et al., 1994; Luce & Pisoni, 1998; McClelland & Elman, 1986; Rodd, 2004), the robust knowledge participants showed in the present studies suggests that encountering a related meaning of a word need not weaken—let alone eliminate—a distinct meaning of the same word in memory. In the network perspective, when an additional meaning is encountered that is related to an existing meaning of a word, a new, if tentative and fragile, meaning can be added. And by use throughout a linguistic community, the word and its additional meaning(s) become conventionalized, changing the meaning from a context-driven inference to something closer to a stored exemplar. Speakers will learn to use the term in appropriate scenarios by means of a network of related, conventionalized meanings for each word form (Copestake & Briscoe, 1995; Diessel,

2019; Goldberg, 2019; Lakoff, 1987; Tuggy, 1993). These meanings share attributes with one another on varying dimensions, with both conventions, productive rules, or emergent generalizations creating overlap and connecting the stored instances of use in a network of interrelated senses. In this approach, there may be a central prototype that emerges from the cluster of aggregated experiences and their contexts, but is not an abstract summary representation. Instead, I propose it resembles a mental representation whose rich set of attributes have been strengthened by the surrounding cluster of related meanings. This proposal makes a clear and intuitive prediction that meanings of a common form which share attributes should be learned more quickly (Floyd & Goldberg, 2020) and processed more quickly than homonymous meanings (Rodd, 2004).

The additive nature of the network perspective also avoids several pitfalls of non usage-based word learning. For instance, simply applying a pattern such as "metaphor", without storing past experiences of usage would leave the listener without a sense for which features to observe in the new meaning. An emphasis on additive learning in response to language input can also better account for the origins of word learning in very young children, as abstraction-based accounts do not explain how rules such as "animal-for-meat" or "metaphor" are acquired in the first place. And, moreover, rule-based approaches de-emphasize how experiences of language usage can be stored in memory, ignoring our rich knowledge of the circumstances and context during learning. For example, if we failed to encode the socio-cultural context in which we heard the term "bachelor" being applied, we would have enormous difficulty using and understanding its meaning and connotations (i.e. that it can be used to describe many unmarried men, but not the pope)(Fillmore, 1992a).

Therefore, I further propose that meaning representations are not limited to a single item or exemplar, but reflect the context of usage. This context is learned across many instances and is stored, allowing a word and its associated meanings to relate to background knowledge which is critical for language users to assign the term to a conventionalized meaning. For example, English may extend the adjective *intense* to both a strong taste ("The spice was intense!") and to a strong work ethic ("She's so intense she clocked 80 hours!"). But critically, we remember the contexts in which these extensions are licensed, and are unlikely to extend the meaning to physical strength (the statement "muscle intensity is an important aspect of overall health" would seem odd). The contexts which have become associated with a word, which can be represented as semantic frames, connect each concept to associated information, experiences, and other concepts (Fillmore, 2008; Fillmore & Langẹndoen, 1971). The approach in frame semantics proposes that the word also evokes the the experiences and concepts to which the word is connected.

The the network approach shares much with the frame semantics approach, as it also aims to address the challenge which multiple senses pose. In allowing the multiple meanings of words to be related to one another in a common network or frame, such approaches side-step the challenge of synthesizing a single, "core" definition. In fact, it avoids the problem of requiring any identifiable attribute to be shared across related senses, by simply allowing the senses to be connected.

7.3 Possible implementations

We know of no current models of human word learning that are able to implement a proposal such as the network perspective described above, but it is likely that the network per-

spective can be implemented by a number of learning models. In this section I will argue that a hierarchical Bayesian approach—specifically a nonparametric one—is likely to be particularly suited to model the distributed, generalization-rich additive learning process which incorporates surrounding context. Four criteria are needed to implement the kind of learning described in network perspective: 1) a representation of uncertainty, 2) a way to update uncertainty, 3) generalization, and 4) flexibility over the number of words learned. These requirements can be fulfilled by a probabilistic model (which represents uncertainty) Bayesian inference (to update of uncertainty), hierarchical structure (to generalize meanings across contexts), and a nonparametric approach (to allow for flexibility in the lexicon size acros the lifespan). After describing each of these criteria, I will make clear how the three main ingredients to the network perspective—distributed representations, generalization, and additive learning are carried out by such an approach.

The first ingredient in a successful implementation of the network perspective is the ability to represent uncertainty. This is simply because, during the process of learning, it is frequently unclear what the referent of a new word is, or which additional meaning an old word might refer to (Quine, 1960). Certain prominent prior approaches have failed to represent uncertainty in learning, by positing only one individual meaning for a word with complete confidence, and entirely deleting it if the word is used without the hypothesized meaning present (Gleitman, 1990; Trueswell et al., 2013a). Other models, namely cross-situation word learning models, have implemented something closer to uncertainty, as they track association strengths between words and meanings across all instances (Alishahi et al., 2012; Fazly et al., 2010; Smith & Yu, 2008). As a result, cross-situation models are more robust to learning naturalistic, polysemous meanings (see Chapter 4). For the purpose of

using Bayesian inference (see below), representing the assignment of meanings to words as probability distributions is an ideal way to do this. This allows the model to make use of any representation type, including those from distributed semantics (dimensionality-reduced vectors based in co-occurrence). Probabilistic belief representations have successfully been used to model human learning in a number of domains, including word learning and generalization (Xu & Tenenbaum, 2007), category learning (Tenenbaum, 2000), and the prediction of everyday phenomena (Griffiths & Tenenbaum, 2006). This probabilistic foundation will also facilitate generalization across meanings and contexts through hierarchical Bayesian inference, allowing the model to learn additively by representing uncertainty across meanings witnessed in the same context (local ambiguity), as well as separately representing a different uncertainty of meanings across contexts (global ambiguity).

During naturalistic word learning, it is critical that a model be able to update prior, uncertain knowledge on the basis of new information. This is especially critical for any implementation of the network perspective, as learning is additive, meaning that updating is critical. The process of updating the probabilities between words and meanings can be implemented through Bayesian inference, which offers a principled basis for belief change. This is a relatively simple procedure in which for each hypothesis, in this case the assignment of a label to a meaning, can be calculated as the product of prior information about that hypothesis and a likelihood of the new data being generated by that hypothesis (e.g., how likely it is for that cluster to produce such an meaning). Much like probabilistic belief representations, Bayesian belief updates have been shown to mirror human behavior in many domains of cognition (Griffiths & Tenenbaum, 2006; Tenenbaum, 2000; Xu & Tenenbaum, 2007). The approach of formalizing beliefs as probabilities for a hypothesis

is also useful for generalization, as the posterior on the new meaning can then be used to determine how likely it is that learners should generalize the label. And, because hypotheses in Bayesian inference can represent anything, such as similarity between high-dimensional distributed meanings (see Chapter 6), this approach is flexible to the distributed meaning representations in the network perspective.

Though the basis for generalization of words across meanings was explored in Chapter 6, this investigation did not address a fundamental challenge in the levels of generalization necessary to implement a proposal like the network perspective. Future work will need to address this limitation. Specifically, it is likely to be necessary to represent uncertainty about a mapping between a word and meaning within a context, *e.g.* the mapping between "cap" and baseball cap, which should be stronger than "cap" is with other nearby toys. Equally important, it is necessary to represent uncertainty across contexts, *e.g.* generalizing aspects of the *cap* meaning to other kinds of similar caps in other contexts, such as bottle caps. Within a given context, meanings are likely to compete more for a form than across contexts: evidence for a baseball cap meaning of *cap* in one context is not evidence against a bottle cap meaning in another context. This challenge is far beyond the scope of every existing word learning model in the developmental literature. I suggest that a hierarchical model can succeed in modeling these two different levels of uncertainty: that words should generalize to a different degree within contexts vs. across contexts. This kind of model can be understood relatively simply as a combination of two models. The first can determine the parameters of how to successfully generalize (or constrain) a word's meanings across candidate meanings within a given context, in which they may compete. Then, the second can learn how best to generalize a word across contexts, given information about the various

meanings themselves. This second, higher-level model can also account for how abstract information, such as productive extension patterns or rules, can themselves be learned. Such models have been used in many types of abstract learning tasks (see Tenenbaum et al. 2006 for a review).

Finally, the network perspective requires flexibility in order to learn new words and senses throughout the lifespan. It is not possible to pre-determine the number of clusters or meanings in advance. A nonparametric model is needed to represent uncertainty over the size of the lexicon (Hjort et al., 2010). In some ways, a nonparametric model is like other models that cluster data in that it aims to estimate the best parameters to fit the data. However, it deviates in that it does not assume there is a final size of clusters, instead estimating this from the data itself, and allowing for the possibility that future clusters will develop given additional data (in the same way that adults continue to learn new words through life). Nonparametric models take a different approache than positing a set number of clusters then simplifying through comparison. Instead, these approaches create a process which aims to generate the observed data, and the number of mixture components/hidden variables grows with the data through inference. This approach allows for growth in vocabulary size (number of clusters), and even for the possibility that low-frequency or obsolete words may stop being used, while another word may undergoe semantic drift to take up more meanings. Critically for the network perspective, nonparametric models allow for flexibility in the complexity of a words' meanings; given more context, in the form of data, meanings of words can become more or less distinct from one another, and an association between a candidate meaning and an entirely different meaning can become stronger. This is critical in reflecting the ever-growing nature of our representations. For instance, an indi-

vidual such as a footwear designer may use the term *shoe* to refer only to sneakers, whereas someone without the same contextual input would just as happily use the term to refer to boots, flats, sandals, and mules. In past work, nonparametric models have been successfully used to model many similar behaviors, such as categorization (Gershman & Blei, 2012; Griffiths et al., 2007; Heller et al., 2009; Sanborn et al., 2010), so they may hold promise for complex networks of meaning as well.

7.4 Summary

The network perspective builds on existing approaches in cognitive linguistics and usage-based proposals to account for word learning. It predicts that learners' flexible capacity to use an existing meaning to infer a new extended meaning of a word may be crucial to vocabulary development; older children, adolescents, and even adults must keep pace with extensions for words and meanings found in technical language or even slang. Human vocabulary learning requires a flexible capacity to infer additional meanings, which are then added to a network of knowledge of language. Distributed representations provide the foundation for meanings to relate to one another, generalizations based on use enable learners to quickly observe relatedness, and additive learning allows learners to build a meaning network based on these relationships among meanings.

7.5 To the reader

It is likely not a surprise to hear that this dissertation is inspired by the richness, complexity, and cleverness of words. In what other domain could someone delight that people across the world call "slippers" *flounders*, or "battleships" *armadillos*? In this dissertation, I have

strived to provide the theoretical and empirical basis to argue for a new approach to word learning. So, if the reader has made it this far, I do hope they are compelled by the argument. But, to be honest, I'd be just as happy if the reader experienced even an ounce of the awe that I did when I first realized that language is magic.

References

Alishahi, A., Fazly, A., Koehne, J., & Crocker, M. (2012). Sentence-based attentional mechanisms in word learning: Evidence from a computational model. *Frontiers in Psychology*, 3(200), 1–16.

Apostle, H. G. (1980). *Aristotle's Categories and Propositions*.

Aravind, A., Villiers, J., Pace, A., Valentine, H., Golinkoff, R., Hirsh-Pasek, K., & Sweig Wilson, M. (2018). Fast mapping word meanings across trials: Young children forget all but their first guess. *Cognition*, 177, 177–188.

Aristotle (BCE). Categories. *The Complete Works of Aristotle, 2 vols (One-Volume Digital Edition)*.

Armstrong, B. & Plaut, D. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Com- putational and behavioral evidence. In B. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 273–278). Austin, TX: Cognitive Science Society.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13(3), 263–308.

Azuma, T. & Van Orden, G. (1997). Why safe is better than fast: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, 36, 484–504.

Backscheider, A. & Gelman, S. (1995). Children's understanding of homonyms. *Journal of Child Language*, 22, 107–127.

Baddeley, A., Eysenck, M., & Anderson, M. (2009). *Memory*. New York, NY: Psychology Press.

Baldwin, D. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, 20(2), 395–418.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278.

Barracuda1983 (2006). File:pipistrellus flight2. Retrieved from.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2019). Capturing human categorization of natural images at scale by combining deep networks and cognitive models. *arXiv preprint arXiv:1904.12690*.

Bracken, J., Degani, T., Eddington, C., & Tokowicz, N. (2017). Translation semantic variability: How semantic relatedness affects learn- ing of translation-ambiguous words. *Bilingualism: Language and Cog- nition*, 20, 783–794.

Bréal, M. (1897). Une science nouvelle: la sémantique. *Revue des Deux Mondes (1829-1971)*, 141(4), 807–836.

Britton, B. (1978). Lexical ambiguity of words used in english texts. *Behavior Research Methods & Instrumentation*, 10(1), 1–7.

Brocher, A., Koenig, J., Mauner, G., & Foraker, S. (2017). About sharing and commitment: The retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience*, 33, 443–466.

Brosnan, M., Scott, F., Fox, S., & Pye, J. (2004). Gestalt processing in autism: Failure to process perceptual relationships and the implications for contextual understanding. *Journal of Child Psychology and Psychiatry and Allied Disciplines*.

Caramazza, A. & Grober, E. (1976). Polysemy and the structure of the subjective lexicon. *Georgetown University roundtable on languages and linguistics. Semantics: Theory and application*, (pp. 181–206).

Casenhiser, D. (2005). Children's resistance to homonymy: An exper- imental study of pseudohomonyms. *Journal of Child Language*, 32, 319–343.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1, 33.

Copestake, A. & Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of semantics*, 12(1), 15–67.

Crisco (2014). File:cap of "aqua"-brand water bottle. Retrieved from.

Cunningham, A. (2005). Vocabulary growth through independent read- ing and reading aloud to children. In E. Hiebert & M. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 45–68). Mahwah, NJ: Erlbaum.

Dautriche, I. (2015a). *Weaving an ambiguous lexicon*. PhD thesis, Sorbonne Paris Cité.

Dautriche, I. (2015b). Weaving an ambiguous lexicon.

Dautriche, I., Fibla, L., Fievet, A.-C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive psychology*, 104, 83–105.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (2016). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, (pp. 1–21).

de Villiers, J., Myers, B., & Stainton, R. (2013). Revisiting pragmatic abilities in autism spectrum disorders. *Pragmatics & Cognition*, 21(2), 253–269.

Delgado, C. (2012). File:bic blue pen cap. Retrieved from.

Dell, G. & Gordon, J. (2003). Neighbors in the lexicon: Friends or foes. In N. Schiller & A. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 9–37). New York, NY: Mouton.

Deng, W. & Sloutsky, V. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, 91, 24–62.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Diessel, H. (2019). *The grammar network*. Cambridge University Press.

Durkin, K. & Manning, J. (1989a). Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, 18(6), 577–612.

Durkin, K. & Manning, J. (1989b). Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, 18, 577–612.

Egstrom, G., Weltman, G., Baddeley, A., Cuccaro, W., & Willis, M. (1972). Underwater work performance and work tolerance. *Report*, (51).

Eigsti, I. & Schuh, J. (2016). Language acquisition in asd: Beyond standardized language measures. In *Innovative Investigations of Language in Autism Spectrum Disorder*.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4), 547–582.

Falkum, I. (2015). The how and why of polysemy: A pragmatic account. *Lingua*, 157, 83–99.

Fauconnier, G. (1994). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge University Press.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computa- tional model of cross-situational word learning. *Cognitive Science*, 34, 1017–1063.

Fellbaum, C., Ed. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fernald, A., Marchman, V., & Weisleder, A. (2013). Ses differences in language processing skill and vocabulary are evident at 18 months. *Developmental science*, 16(2), 234–248.

Fillmore, C. (1992a). : (pp. 35–60). Berlin: Moulton de Gruyter.

Fillmore, C. & Atkins, B. (2000). Describing polysemy: The case of 'crawl. In *Polysemy: Theoretical and Computational Approaches*.

Fillmore, C. & Langendoen, D. (1971). *Studies in linguistic seman- tics*. New York, NY: Holt, Rinehart & Winston.

Fillmore, C. J. (1992b). Corpus linguistics" or "computer-aided armchair linguistics". volume 82 (pp. 35–60).

Fillmore, C. J. (2008). Frame semantics. In *Cognitive linguistics: Basic readings* (pp. 373–400). De Gruyter Mouton.

Fillmore, C. J. & Langėndoen, D. T. (1971). Studies in linguistic semantics.

Flavell, J., Speer, J., Green, F., August, D., & Whitehurst, G. (1981). The development of comprehension monitoring and knowledge about communication. *Monographs of the Society for Research in Child Development*, 46, 1–65.

Floyd, S., Dalawella, K., Lew-Williams, C., Goldberg, A. E., & Griffiths, T. (2021a). Modeling rules and similarity in colexification. In *To appear in the Proceedings of the 43rdd Annual Conference of the Cognitive Science Society*.

Floyd, S. & Goldberg, A. E. (2020). Children make use of relationships across meanings in word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Floyd, S., Goldberg, A. E., & Lew-Williams, C. (2020). Toddlers recognize multiple meanings of polysemous words. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

Floyd, S., Jeppsen, C., & Goldberg, A. E. (2021b). Brief report: Children on the autism spectrum are challenged by complex word meanings. *Journal of Autism and Developmental Disorders*, 51(7), 2543–2549.

François, A. (2008). Semantic maps and the typology of colexification. *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 106, 163.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, 20(5), 578–585.

Frazier, L. & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, 29, 181–200.

Geeraerts, D. & Gentner, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, 4, 223–272.

Gershman, S. J. & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1), 1–12.

Gleitman, L. (1990). The structural sources of verbs meanings. *Language*.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language*.

Godden, D. & Baddeley, A. (1975). Context dependent memory in two natural environments. *Acquisition*, 1, 3–55.

Goldberg, A. (2019). *Explain me this: Creativity, competition and the partial productivity of constructions*. Princeton, NJ: Princeton Univer.

Golinkoff, R., Hirsh-Pasek, K., Bailey, L., & Wenger, N. (1992).

Gombert, J. (1992). *Metalinguistic development*. Chicago, IL: Univer- sity of Chicago Press.

Gordon, P. (2004). Numerical cognition without words: Evidence from amazonia. *Science*, 306(5695), 496–499.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, 114(2), 211.

Griffiths, T. L. & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, 17(9), 767–773.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of verbal learning and verbal behavior*, 18(4), 441–461.

Harmon, Z. & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology*, 98, 22–44.

Hartley, C., Trainer, A., & Allen, M. (2019). Investigating the relationship between language and picture understanding in children with autism spectrum disorder. *Autism*.

Hawa, V. & Spanoudis, G. (2014). Toddlers with delayed expressive language: An overview of the characteristics, risk factors and language outcomes. *Research in developmental disabilities*, 35(2), 400–407.

Heine, B. (1992). Grammaticalization chains. *Studies in Language*, 16, 335–368.

Heine, B. & Reh, M. (1984a). *Grammaticalization and reanalysis in African languages*. Buske, Hamburg.

Heine, B. & Reh, M. (1984b). *Grammaticalization and reanalysis in African languages*. Hamburg: Helmut Buske Verlag.

Heller, K., Teh, Y. W., & Gorur, D. (2009). Infinite hierarchical hidden markov models. In *Artificial Intelligence and Statistics* (pp. 224–231).: PMLR.

Henry, L., Farmer, C., Manwaring, S., Swineford, L., & Thurm, A. (2018). Trajectories of cognitive development in toddlers with language delays. Research in Developmental Disabilities.

Hinton, G. & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98, 74–95.

Hintzman, D. L. (1986). " schema abstraction" in a multiple-trace memory model. *Psychological review*, 93(4), 411.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press.

Horst, J. & Hout, M. (2016). The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, 48, 1393–1409.

Horst, J. & Samuelson, L. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, 13, 128–157.

Hume, D. (1748/1955). An inquiry concerning human understanding with a supplement an abstract of a.

Jackendoff, R. (1988). Semantics and cognition. second printing 1985).

Johnson, M., Boyd, J., & Goldberg, A. (2012). Construction learning in children with autism. In *Proceedings of the Thirty-Sixth Annual Boston University Conference on Language Development: Supplemental* (pp. 1–15).

Joordens, S. & Besner, D. (1994). When banking on meaning is not (yet) money in the bank: Explorations in connectionist modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1051– 1062.

Katz, J. & Postal, P. (1964). An integrated theory of linguistic descriptions (book review). *Integrated Theory of Linguistic Descriptions*, 34, 670–674.

Kawamoto, A. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, 32, 474–516.

Kawamoto, A., Farrar, W., & Kello, C. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Per- ception and Performance*, 20, 1233–1247.

Kim, S., Paul, R., Tager-Flusberg, H., & Lord, C. (2014). Language and communication in autism. In *Handbook of Autism and Pervasive Developmental Disorders*. Fourth edition.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., et al. (2016). D-place: A global database of cultural, linguistic and environmental diversity. *PloS one*, 11(7), e0158391.

Klein, D. & Murphy, G. (2001). The representation of polysemous words. *Journal of Memory and Language*, 45, 259–282.

Klein, D. & Murphy, G. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, 47, 548 –570.

Klepousniotou, E. & Baum, S. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, 20, 1–24.

Klinger, L. & Dawson, G. (2001). Prototype formation in autism. *Development and Psychopathology*, 13(1), 111–124.

Kuhl, P., Coffey-Corina, S., Padden, D., & Dawson, G. (2005). Links between social and linguistic processing of speech in preschool children with autism: Behavioral and electrophysiological measures. *Developmental Science*.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4), 978–990.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmertest pack- age: Tests in linear mixed effects models. *Journal of Statistical Soft- ware*, 82, 1–26.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.

Lakoff, G. & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3(3), 299–321.

Landau, B., Smith, L., & Jones, S. (1998). Object shape, object function, and object name. *Journal of Memory and Language*, 38, 1–27.

Langacker, R. (1987). Foundations of cognitive grammar. In *Theoretical Prerequisites*, volume 1. Stanford, CA: Stanford University Press.

Lehrer, A. (1990a). Polysemy, conventionality, and the structure of the lexicon. *Cognitive Linguistics*, 1, 207–246.

Lehrer, A. (1990b). Polysemy, conventionality, and the structure of the lexicon. *Cognitive linguistics*, 1(2), 207–246.

Lewandowska-Tomaszczyk, B. (2012). Polysemy, prototypes, and radial categories. In *The Oxford Handbook of Cognitive Linguistics*.

Litman, L., Robinson, J., & Abberbock, T. (2017a). Turkprime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, 49(2), 433–442.

Litman, L., Robinson, J., & Abberbock, T. (2017b). Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49, 433–442.

Luce, P. & Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1–36.

Lukyanenko, C. & Fisher, C. (2016). Where are the cookies? two- and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition*, 146, 349–370.

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, 23(20), 2023–2027.

MacWhinney, B. (2000). *The CHILDES Project: Transcription format and programs*, volume 1. IAP.

Malt, B. (2010). Naming artifacts: Patterns and processes. In B. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 1–38). Amsterdam, the Netherlands: Elsevier.

Markman, E. & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121–157.

McClelland, J. & Elman, J. (1986). The trace model of speech perception. *Cognitive Psychology*, 18, 1– 86.

McCloskey, M. E. & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, 6(4), 462–472.

McKinley, S. C. & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, 21(1), 128.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow.

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, 85(3), 207.

Medin, D. L. & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, 7(4), 241.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011).

Merriman, W., Bowman, L., & MacWhinney, B. (1989). The mutual.

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the psychonomic society*, 7(3), 283–284.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Minshew, N., Goldstein, G., Muenz, L., & Payton, J. (1992). Neuropsychological functioning in nonmentally retarded autistic individuals. *Journal of Clinical and Experimental Neuropsychology*.

Molesworth, C., Bowler, D., & Hampton, J. (2005). The prototype effect in recognition memory: Intact in autism? *Journal of Child Psychology and Psychiatry and Allied Disciplines*.

Mottron, L., Dawson, M., Soulières, I., Hubert, B., & Burack, J. (2006). Enhanced perceptual functioning in autism: An update, and eight principles of autistic perception. *Journal of Autism and Developmental Disorders*.

Mottron, L. & Morasse, K. (2001). A study of memory functioning in individuals with autism. *J. Child Psychol. Psychiat. Association for Child Psychology and Psychiatry*, 42(2), 253–260.

Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and Developmental Disorders*.

Murphy, G. (2004a). *The big book of concepts*. MIT press.

Murphy, G. (2004b). *The big book of concepts*. Cambridge, MA: MIT Press.

Navigli, R. & Ponzetto, S. (2012a). Babelnet: The automatic construc- tion, evaluation and application of a wide-coverage multilingual seman- tic network. *Artificial Intelligence*, 193, 217–250.

Navigli, R. & Ponzetto, S. P. (2012b). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193, 217–250.

Nosofsky, R., Gluck, M., Palmeri, T., McKinley, S., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961. *Memory & Cognition*, 22, 352–369.

Nunberg, G. (1979a). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, 3, 143–184.

Nunberg, G. (1979b). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and philosophy*, 3(2), 143–184.

Ostler, N. & Atkins, B. T. S. (1991). Predictable meaning shift: Some linguistic properties of lexical implication rules. In *Workshop of SIGLEX (Special Interest Group within ACL on the Lexicon)* (pp. 87–100).: Springer.

Ozonoff, S. & McEvoy, R. (1994). A longitudinal study of executive function and theory of mind development in autism. *Development and psychopathology*, 6(3), 415–431.

Paul, R. (2003). Promoting social communication in high functioning individuals with autistic spectrum disorders. *Child and Adolescent Psychiatric Clinics of North America*, 12(1), 87–106.

Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122, 280–291.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. New York, NY: Viking.

Plaisted, K., O'Riordan, M., & Baron-Cohen, S. (1998). Enhanced visual search for a conjunctive target in autism: A research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines*.

Plebanek, D. & Sloutsky, V. (2017). Costs of selective attention: When children notice what adults miss. *Psychological Science*, 28, 723–732.

Potrzeba, E., Fein, D., & Naigles, L. (2015). Investigating the shape bias in typically developing children and children with autism spectrum disorders. *Frontiers in Psychology*.

Pustejovsky, J. (1998). *The generative lexicon*. MIT press.

Pye, C. (2017). *The comparative method of language acquisition research*. Chicago, IL: University of Chicago Press.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Rabagliati, H., Marcus, G. F., & Pylkkänen, L. (2010). Shifting senses in lexical semantic development. *Cognition*, 117(1), 17–37.

Rabagliati, H. & Snedeker, J. (2013). The truth about chickens and bats.

Ramiro, C., Srinivasan, M., Malt, B., & Xu, Y. (2018a). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences of the United States of America*, 24, 1354–1360.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018b). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, 115(10), 2323–2328.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The.

Reimers, N. & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of verbal learning and verbal behavior*, 12(1), 1–20.

Rodd, J. (2004). When do leotards get their spots? semantic activation of lexical neighbors in visual word recognition. *Psychonomic Bulletin & Review*, 11, 434–439.

Rodd, J., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M., & Davis, M. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition*, 40, 1095–1108.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002a). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2), 245–266.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002b). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46, 245–266.

Rosch, E. (1978). Principles of categorization. *Concepts: core readings*, 189.

Rosch, E. & Mervis, C. (1975). Family resemblances: Studies in the.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural.

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology*, 22(4), 460–492.

RStudio (2017). *Integrated development for R*. Boston, MA: RStudio.

Rumelhart, D. E. & Ortony, A. (1977). The representation of knowledge in memory. anderson rc, spiro rj, montague we (eds) schooling and the acquiation of knowledge.

Rutherford, M. & McIntosh, D. (2007). Rules versus prototype matching: strategies of perception of emotional facial expressions in the autism spectrum. *Journal of Autism and Developmental Disorders*, 37(2), 187–196.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., et al. (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1), 1–12.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, 117(4), 1144.

Schmitt, N. & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge university press.

Schreibman, L. (2000). Intensive behavioral/psychoeducational treatments for autism: Research needs and future directions. *Journal of Autism and Developmental Disorders*.

Schuh, J., Eigsti, I., & Mirman, D. (2016). Discourse comprehension in autism spectrum disorder: Effects of working memory load and common ground. *Autism Research*.

Siedlecki, P. (2018). Image of baseball bat. Retrieved from.

Sinha, P., Kjelgaard, M., Gandhi, T., Tsourides, K., Cardinaux, A., Pantazis, D., & Held, R. (2014). Autism as a disorder of prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 111(42), 15220–15225.

Smith, J. & Minda, J. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1411–1436.

Smith, L. & Kemler, D. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual de- velopment. *Journal of Experimental Child Psychology*, 24, 279–298.

Smith, L. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological science*, 13(1), 13–19.

Smoke, K. L. (1932). An objective study of concept formation. *Psychological Monographs*, 42(4), i.

Srinivasan, M., Al-Mughairy, S., Foushee, R., & Barner, D. (2017). Learn- ing language from within: Children use semantic generalizations to infer word meanings. *Cognition*, 159, 11–24.

Srinivasan, M., Berner, C., & Rabagliati, H. (2019). Children use polysemy to structure new word meanings. *Journal of Experimental Psychology: General*, 148, 926–942.

Srinivasan, M. & Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157, 124–152.

Srinivasan, M. & Snedeker, J. (2011). Judging a book by its cover and its contents: The representation of polysemous and homophonous meanings in four-year-old children. *Cognitive psychology*, 62(4), 245–272.

Srinivasan, M. & Snedeker, J. (2014). Polysemy and the taxonomic constraint: Children's representation of words that label multiple kinds. *Language Learning and Development*, 10, 97–128.

Stevens, J., Gleitman, L., Trueswell, J., & Yang, C. (2016). The Pursuit of Word Meanings, Cognitive Science.

Stevens, J., Gleitman, L., Trueswell, J., & Yang, C. (2017a). The pursuit of word meanings. *Cognitive Science*, 41, 638–676.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017b). The pursuit of word meanings. *Cognitive science*, 41, 638–676.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure (Cambridge studies in linguistics*. Cambridge: Cambridge University Press.

Swinney, D. (1979). Lexical access during sentence comprehension: (re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18, 645–659.

Tabossi, P., Colombo, L., & Job, R. (1987). Accessing lexical ambiguity: Effects of context and dominance. *Psychological Research*, 49, 161– 167.

Tek, S., Jaffery, G., Fein, D., & Naigles, L. (2008). Do children with autism spectrum disorders show a shape bias in word learning? *Autism Research*.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. *Advances in neural information processing systems*, 12, 59–65.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7), 309–318.

TexasRebel (2007). File:baseball cap. Retrieved from.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10), 1029–1038.

Tomasello, M. (2009). *Constructing a language*. Harvard university press.

Tovar, ◆., Rodríguez-Granados, A., & Arias-Trejo, N. (2019). Atypical shape bias and categorization in autism: Evidence from children and computational simulations. *Developmental Science*.

Traugott, E. (2014). Toward a constructional framework for research on language change. *Cognitive Linguistic Studies*, 1(1), 3–21.

Trueswell, J., Medina, T., Hafri, A., & Gleitman, L. (2013a). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66, 126 –156.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013b). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66(1), 126–156.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Lin- guistics*, 4, 273–290.

Voeten, C. (2020). buildmer: Stepwise elimination and term reordering for mixed-effects regression. r package version 1.5.

Wiseman, R. (2015). Interpreting ancient social organization: conceptual metaphors and image schemas. *Time and Mind*, 8(2), 159–190.

Wittgenstein, L. (1953). Philosophical investigations.

Xu, F. & Tenenbaum, J. (2007). Word learning as bayesian inference. *Psychological review*, 114(2), 245.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, 201, 104280.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language*.

Yu, C. & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18, 414–420.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, 37(5), 891–921.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: a reassessment.

Zipf, G. (1945a). The meaning-frequency relationship of words. *Journal of General Psychology*, 33, 251–256.

Zipf, G. K. (1945b). The meaning-frequency relationship of words. *The Journal of general psychology*, 33(2), 251–256.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.