Title

Analogical Reasoning Involving Semantic Relations in Children and Machines

Author

Ionescu, Amalia

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Analogical Reasoning Involving Semantic Relations in Children and Machines

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Amalia Ionescu

2024

ABSTRACT OF THE DISSERTATION

Analogical Reasoning Involving Semantic Relations in Children and Machines

by

Amalia Ionescu

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2024

Professor Catherine M. Sandhofer, Co-Chair

Professor Hongjing Lu, Co-Chair

Abstract: Understanding abstract relations is a crucial aspect of human cognition. This thesis explores how children learn to reason about semantic relations, such as antonyms and synonyms, and the factors that enhance their ability to use these relations for analogical reasoning (e.g., understanding that big:small hold the same relation as clean:dirty). Chapter I investigates the early stages of how young children generate antonyms, identifies patterns in their responses across different parts of speech, and compares their abilities to those of relational and natural language models. Chapters II and III focus on how both children and models utilize antonyms to solve verbal analogies, and how relational language cues and

semantic distance affect this skill. This thesis provides a comprehensive examination of the development of antonymy understanding and identifies key factors that can enhance the ability of both children and models to solve verbal analogies.

Keywords: semantic relations, analogical reasoning, semantic distance, computational models

The dissertation of Amalia Ionescu is approved.

Keith Holyoak

Idan Blank

Hongjing Lu, Committee Co-Chair

Catherine M. Sandhofer, Committee Co-Chair

University of California, Los Angeles

2024

Table of Contents

List of Figures

List of Tables

Preface

The seed for this dissertation was planted during my second year of graduate school, in Dr. Idan Blank's lab meeting. Up until that point, I vacillated between various topics, unsure of how I would characterize my research interests when asked. My goal was always to create an interdisciplinary body of work, in collaboration with brilliant minds from different fields. At the time, I did not realize how much those I had already worked with permanently shaped my research interests.

My first foray into research began in Dr. Andrei Cimpian's lab at University of Illinois at Urbana-Champaign. There, I learned the essentials of conducting research with children and became enthralled by the incredible graduate students I had the pleasure to work with: Drs. Christina Tworek, Zach Horne, Lin Bian, Larisa Hussak, Shelby Sutherland, and Daniel Storage. I am especially thankful for Christina, who mentored me on my senior thesis project with great care and attention, and to Zach, who encouraged me in the most Zach way to apply to UCLA and who always provided candid and essential feedback. I am most thankful for Andrei, who has been one of my greatest supporters throughout my academic journey. I feel so endlessly grateful and lucky to have had his mentorship as an undergraduate and beyond, and I feel certain that without his help I would have never gotten where I am now.

Though I knew I wanted to pursue research after my undergraduate degree, I knew that I needed more experience before I could begin a PhD. Through Andrei's recommendation, I spent two formative years in Dr. Dedre Gentner's lab at Northwestern University as her lab coordinator. I recall so distinctly being in constant awe of everybody in that lab: Drs. Nina Simms, Francisco Maravilla, Christian Hoyos, Ruxue Shao, Kensy Cooperrider. I felt as though I

had snuck into an intellectual powerhouse where I was to try to absorb as much knowledge as I could. Extraordinarily, their intellect was matched only by their immense kindness and generosity. I will always remember the lunch conversations, the holiday parties, and the many laughs. I am so lucky to have gotten to work with Dedre – a giant in the field, and a wonderful, patient, funny, endlessly impressive advisor.

When I started my PhD, I felt drawn to the idea of incorporating culture into the study of cognitive development. My first project in graduate school was under the mentorship of Dr. Patricia Greenfield. Patricia encouraged me to think deeply about the effects of our environments on cognition, and to not let my research questions be constrained by resources. Indeed, as impossible as my first study seemed at the time, with her help and determination we were able to find collaborators and resources to help us finish it. I will always have deep admiration for Patricia.

As my research interests continued to evolve, I began working more closely with Dr. Catherine Sandhofer, who remained my primary advisor throughout my graduate career. During our many meetings over the years, Cathy has been ever supportive and encouraging of everything I have tried to do, and always helped me persist with my research and writing. I am lucky to have benefited from her expertise and insights on language and cognitive development research, the freedom and support to pursue the ideas I had, and to have been always treated with so much kindness and understanding. I cannot begin to imagine what my graduate career would have been like without Dr. Sandhofer's mentorship.

During my second and third years, I began collaborating with Drs. Hongjing Lu, Keith Holyoak, and Idan Blank. I believe now that most of my ideas, especially those that became my

dissertation, emerged during our many lab meetings, which were both intellectually cultivating and wonderfully humorous. As my interests gradually began shifting back to analogical reasoning, I was lucky to have Keith as a mentor and collaborator. My deep interest in this topic will never cease to exist, and that is in large part because of Keith and Dedre. Keith's wisdom and incredible insight into essentially every topic will always have my admiration. I am thankful to Hongjing for her mentorship, encouragement, and enthusiasm as I began my first foray into computational work. Every interaction left me optimistic and intellectually invigorated, with a great desire to learn more. Finally, I have so much gratitude for Idan, for his immense kindness and generosity, and his incredible ability to think deeply about any idea we discussed. To have had such a wonderful group of mentors during my graduate career has been an honor.

Apart from my mentors, my years in graduate school have been indelibly marked by Hunter Priniski, Adriana Mendez Leal, Renée Zhu, Nick Ichien, Mason McClay, Josh MacNeal, Erika Blair, Kelly Medrano, and Beck, who provided me with so much support and care over the years and who made my life an adventure. There is much to be said about our times together, but I owe so much of my happiness to them.

My eternal gratitude to my family, especially my mom, for all the love, time, and support they have given me so that I can chase any dream I had. They have been a driving force behind every opportunity I have had, and the source of constant care. Thank you for everything.

Vita

EDUCATION

University of California, Los Angeles (UCLA), Los Angeles, CA
M.A. Developmental Psychology — December 2019
Minor: Cognitive Psychology

University of Illinois at Urbana-Champaign — May 2016
B.S. in Psychology
Minor: Philosophy

AWARDS, FELLOWSHIPS, & GRANTS

UCLA Psychology Dissertation Fellowship	2023
Diverse Intelligences Summer Institute Fellowship	2022
UCLA Summer Mentored Research Fellowship (SMRF)	2021
Patricia Greenfield International Field Research Award	2020
UCLA Graduate Summer Research Mentorship Award (GSRM)	2019
University of Illinois Distinction in Psychology Award	2016
University of Illinois Dean's List	2015– 2016
Psi Chi International Honor Society in Psychology Member	2015– 2016

PUBLICATIONS

Ionescu, A., Furdui, R., Gavreliuc, A., Greenfield, P.M., & Weinstock, M. (2023). The Effects of Sociocultural Changes on Epistemic Thinking Across Three Generations in Romania. *PLOS One, 18*(3), e0281785. https://doi.org/10.1371/journal.pone.0281785.

Ionescu, A., Lu, H., Holyoak, K.J., & Sandhofer, C.M. (2022). Children's Acquisition of the Concept of Antonym Across Different Lexical Classes. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (pp. 2526-2531). Toronto, Canada: Cognitive Science Society.

CONFERENCE POSTERS & PRESENTATIONS

Ionescu, A., Lu, H., Holyoak, K.J., Sandhofer, C.M. (March, 2023). "*The role of language in preschoolers' understanding of antonyms*" Poster at the 2023 Biennial Meeting for Society for Research in Child Development.

Ionescu, A., Lu, H., Holyoak, K.J., Sandhofer, C.M. (November, 2022). "*Children's Antonym Understanding Across Parts of Speech*" Poster at the Psychonomic Society 63rd Annual Meeting.

Ionescu, A., Lu, H., Holyoak, K.J., Sandhofer, C.M. (July, 2022). "*Children's Acquisition of the Concept of Antonym Across Different Lexical Classes*" Flash talk at the 44th Annual Meeting of the Cognitive Science Society.

Ionescu, A., Furdui, R., Gavreliuc, A., Greenfield, P.M. (2021, November). "*Epistemic Thinking and Social Change: An Inter-Generational Analysis of the Transition to Post-Communism*" Talk at the Association of Psychologists in Romania Conference.

Ionescu, A., Tworek, C.M., Sandhofer, C. & Cimpian, A. (2021, July). *"The Effects of Messages about Intellectual Ability on Children's Activity Preferences"* Poster presented at Cognitive Science Society.

Ionescu, A., Tworek, C.M., Sandhofer, C. & Cimpian, A. (2021, April).*"The Effects of Gender Stereotypes on Children's Activity Choices"* Talk at Society for Research in Child Development.

Ionescu, A., Greenfield, P.M. (2019, December). *"Cross-Generational Differences in Epistemological Development"* Talk at University of California, Los Angeles Developmental Psychology Forum, Los Angeles, CA.

Ionescu, A., Greenfield, P.M. (2019, June). *"The Effects of Sociocultural Change on Epistemic Development"* Talk at Symposium on Cognition and Language Development, Los Angeles, CA.

Ionescu, A., Greenfield, P.M. (2019, May). *"Social Change and Epistemic Thinking"* Talk at the University of California, Los Angeles Developmental Psychology Forum, Los Angeles, CA.

Ionescu, A., Tworek, C.M., & Cimpian, A. (2016, April). *"Feminizing Activities Counteracts the Negative Effects of Gender Stereotypes"* Poster presented at the University of Illinois at Urbana-Champaign Undergraduate Research Symposium, Champaign, IL.

Ionescu, A., Tworek, C.M., & Cimpian, A. (2016, April). *"Feminizing Activities Counteracts the Negative Effects of Gender Stereotypes"* Poster presented at the University of Illinois at Urbana-Champaign Psychology Department Honors and Capstone Research Fair, Champaign, IL.

TEACHING EXPERIENCE

Teaching Fellow, University of California, Los Angeles	
Research Methods in Psychology	Summer 2023, Spring 2024
Choice Architecture (Anderson School of Business)	Winter 2024
Developmental Psychology Lab	Spring 2023
Dynamic Perspectives on Parenting	Winter 2023
Language Development	Fall 2022
Teaching Associate, University of California, Los Angeles	
Developmental Psychology	Winter, Spring 2021
Cognitive Development	Fall 2020, Winter 2022
Cognitive Psychology	Summer 2020
Culture and Human Development	Spring 2020
Teaching Assistant, University of California, Los Angeles	
Research Methods in Psychology	Winter 2020, Fall 2021
Research Methods in Developmental Psychology	Fall 2019
Introductory Psychology	Winter 2019

Introduction

The ability to reason about semantic relations (e.g., synonyms, antonyms, function) is crucial to cognitive development. To do so, children must first learn the meaning of words, how these meanings are related across pairs of words, and finally, how these binary relations can be used to reason by analogy across various instantiations. Such relations are first formally taught to children in elementary school and followed up repeatedly throughout formal schooling, where the ability to complete verbal analogy problems (e.g., rich : poor :: big : small) is tested, including on standardized tests such as the SAT.

Although a large body of research has examined children's ability to learn relational words and categories (Hall & Waxman, 1993; Asmuth & Gentner, 2005; Gentner et al., 2011), less is known about how children learn abstract semantic relations. Considering the need to be able to use such relations to reason by analogy, it is crucial to understand the developmental trajectory of learning abstract semantic relations and the types of support that facilitate this type of learning.

Semantic Relations

We must first examine what semantic relation comprehension entails to understand how the factors involved in word and category learning might affect semantic relation learning. Semantic relations (associations between meanings of words) are considered fundamental components of language and thought. Semantic relations can be compartmentalized into

categories: antonymy, synonymy, function, part-whole relations, cause-purpose, space-time, case relations, etc. (Lu et al., 2019; Landis et al., 1987; Murphy, 2003).

To reason about abstract relations, children must first learn the meaning of the individual words (e.g., understand the meaning of the word "happy"), how these meanings are related across words (e.g., learn that "happy" and "sad" are both emotions and belong to the same category), and the binary relations that pairs of words have with each other (e.g., "happy" and "sad" both convey emotions, but they are also opposites and share a contrast relation). Afterward, learners can identify various instantiations of the same abstract relation. For example, one can grasp that "fly" is related to "bird" in the same way as "cut" is related to "knife" (function) or that "night" is related to "day" in the same way as "poor" is related to "rich" (opposite). Machine-learning models, such as Word2Vec and BART (Bayesian Analogy with Relational Transformations), which should be comparable to adults in semantic relation processing, show variation in their ability to learn specific semantic relations (Lu et al., 2019). Therefore, there is reason to believe that there might also be differences in the development of semantic relation comprehension during early childhood.

Indeed, research has shown differences in children's comprehension of various semantic relations, though the focus has been mostly on synonyms and antonyms (Heidenheimer, 1978; Garnham et al., 2000). Moreover, research shows that children's first understanding of the antonym relation begins with substantially higher accuracy than other relations, such as part-whole inclusion, synonyms, and class inclusion (Landis et al., 1987; Heidenheimer, 1978). Studies find that children's understanding of semantic relations increases between second and eighth grade in two respects: knowledge about what each relation is and is not (Landis et al.,

1987). For example, second graders understand the antonym relation well (i.e., second graders understand that the relationship between two words can be categorized as an antonym if the words entail opposition). Yet, at this age, there are still a number of intrusions into the relation (i.e., children do not yet fully understand that antonyms do not include *any* two words that might overlap in meaning, e.g., salt and sea). However, by eighth grade, children's understanding of the antonym relation is refined in both respects (Landis et al., 1987).

Much of the research on children's understanding of semantic relations has failed to consider the possibility of different elements that require processing (e.g., form class, semantic content, familiarity with the stimulus words) for different types of word stimuli. For example, previous research on antonym understanding has largely focused on adjectives and has failed to examine other types of lexical classes, despite research showing that children acquire different parts of speech starting at different times, beginning with nouns, followed by adjectives and verbs (Nelson, 1973; Sandhofer & Smith, 2007). Similarly, the semantics of each word or pair of words and the role this plays in early semantic relation understanding have not been systematically examined. This is important as the semantic complexity of a word or a pair of words might make the relation either more or less noticeable. Research using word association and false recognition tasks found that first graders produced significantly more antonyms than synonyms when controlling for these factors. However, the production of synonyms does seem to increase with age (Heidenheimer, 1978). Thus, it is clear that children can grasp antonymy from a young age and that antonymy plays an important part in young children's cognitive development (Jones & Murphy, 2005).

Antonymy

The antonym relation involves understanding the concept of "opposite," which makes it possible to identify an indefinite number of instantiations of the same abstract relation (e.g., being able to reason about "rich and poor" sharing the same relation as "ugly and pretty"). Antonyms are a unique semantic relation because the words that share the relation typically belong to the same category or are closely associated (e.g., happy and sad both describe emotions), yet they differ maximally typically on a single dimension (e.g., happy and sad are opposites even though they are both emotions). Empirical research assessing children's understanding of the concept of "opposite" tends to fall into two categories: discourse studies and metalinguistic studies. Discourse studies primarily center on children's spontaneous usage of antonyms with children as young as two years old (Tribushinina et al., 2013; Jones & Murphy, 2005). Furthermore, there is evidence that young children's usage of explicit contrast in speech (e.g., "give me the big piece, not the small piece") is largely associated with parents' tendency to do the same. This is particularly important as reasoning about contrasts might facilitate attention to the various dimensions that antonyms can be evaluated (Tribushinina et al., 2013).

On the other hand, metalinguistic studies evaluate children's ability to work with the metalinguistic vocabulary of opposition. They primarily involve verbal games in which children respond to questions such as "What is the opposite of X?" Other studies of this type used free association tasks, which showed that children tend to respond with a word that is closely associated with the stimulus word (e.g., dark-night) prior to five years of age, while older

children tend to respond with a word that is semantically opposite to the stimulus (e.g., dark-light) (Entwistle et al., 1964). The verbal component of these instruments might explain why metalinguistic studies show that the antonym relation becomes salient to children only around five years of age. Previous research ties relational reasoning with executive functioning. Namely, research on the development of working memory indicates that children's working memory capacity increases with age (Gathercole et al., 2004), and this contributes to their ability to process binary relations (a relation between two arguments/objects) around two years of age and, later, ternary relations (a relation between three arguments/objects) after five years of age (Halford, 1993; Andrew & Halford, 2002). Therefore, it could be that young children's relatively limited working memory capacity contributes to their limited ability to detect relations in these types of verbal tasks. However, when the verbal component of such tasks was eliminated, children understand the "opposite" relation much earlier, specifically, around four years of age (Phillip & Pexman, 2015). Using a non-verbal opposite task, researchers found that labeling the objects and providing a label for the opposite relation helped four- and five-year-old children understand this relation (Phillip & Pexman, 2015).

When considering the extant literature on semantic relation learning, it is clear that it is largely divorced from the literature on traditional word and category learning, which has a history of examining the acquisition process. Specifically, the factors that facilitate word and category learning, especially relational words and categories, have not been focused on as meticulously in semantic relation learning in children. While some work does show that language, namely labels, plays a role in semantic relation learning the way it does in category learning, research on this topic is limited (Phillips & Pexman, 2015). One possibility for this

oversight might be because semantic relations are abstract and are often presented verbally using the words themselves rather than as pictures, which are typically presented with other abstract relational categories. Most studies have examined children's spontaneous production and free association of semantic relations. However, if researchers were to take an experimental approach using non-verbal stimuli, as they often do for relation learning studies, they could implement the same strategies that have been shown to facilitate relational word/category learning, such as language (labeling), context (structural alignment), and comparison. If these factors were considered, children might show more understanding of complex semantic relations such as the "opposite" relation than when spontaneously generating these relations.

Analogical Reasoning

A marker of semantic relation understanding is successfully drawing analogies between pairs of words that hold the same relation. There are two types of analogy problems: the first, referred to as the classical analogy, takes the form A:B::C:D (e.g., Hot:Cold::Small:Big), and the second is the problem solving analogy, in which the solution to one problem is explained and then using analogy, can aid in solving a novel, more difficult problem. Both types of analogies are commonly used in research, but the classical analogy task is the most appropriate for examining semantic knowledge understanding (Lu et al., 2019).

Development of Analogical Reasoning

First, it is crucial to elucidate the developmental trajectory of analogical reasoning – when do children draw analogies across different instances? Traditionally, analogical reasoning has been considered accessible only for older children and adults, as it necessitates relational

knowledge (information about the higher-order relations that the analogy depends on), an ability to make relational inferences (realizing that the link between A and B can be applied to C and D), and knowledge of task requirements (clarity on what the aim of the task is) (Singer-Freeman, 2005). While early work suggested that children are unable to reason analogically until about 13-14 years of age (Piaget et al., 1977), more recent evidence suggests that children as young as age three can successfully complete analogical reasoning tasks as long as they know the relations involved (Goswami & Brown, 1989; Richland et al., 2006). However, increased domain knowledge is not the only factor contributing to children's increased understanding of analogical reasoning across development. One theory is that children undergo a "relational shift," in which they shift their attention from featural similarity to relational similarity (Gentner & Rattermann, 1991). In other words, while young children first attend primarily to perceptual features, with time, they begin attending to relational similarities. This relational shift occurs at different ages for different domains and is contingent upon children's knowledge of that domain. For example, if shown a picture of a dog chasing a cat and another picture of a boy chasing a girl while a cat is present in the background, younger children (aged 3) tend to match the cats in both photos (featural similarity). In contrast, older children (aged 5) tend to match the cat from the first picture to the girl from the second, as both are being chased (relational similarity).

However, there are factors that support children's analogical reasoning, including relational familiarity, relational language, and comparison across instances. Relational similarity in analogical tasks involves using familiar examples, such as physical causality (Goswami & Brown, 1990). If children are given problems involving examples of relations that they

frequently encounter in their daily lives, it is easier for them to solve an analogical reasoning problem. Another factor is awareness of relational language, including labels and names for the relations involved in an analogical reasoning task. Research has shown that using relational labels for objects in a task helps children notice and manipulate memories of relational similarities, similar to how labels help children learn categories (Loewenstein & Gentner, 1998; Rattermann & Gentner, 1998). Lastly, comparison across instances helps children extract higher-order relations in analogical reasoning problems, as it highlights the shared relational structure between examples (Gentner & Namy, 1999; Anggoro et al., 2005).

Semantic Distance

An important factor that should be controlled for when studying how semantic relations can be used to reason by analogy is semantic distance. Previous work has shown that semantic distance could be represented as Euclidean distance in a semantic space. For example, a "robin" is closer in the semantic space to "bird" than it is to "animal." Findings suggest that the statement "a robin is a bird" is verified faster than the statement "a robin is an animal" (Collins & Quillian, 1969). Given that a robin is a subset of a bird, and a bird is a subset of an animal, the representation of a robin and a bird are closer together in an underlying semantic structure than that of a robin and an animal. This subset effect manifests as the increasing time it takes to confirm a semantic relation between two nouns as the semantic distance between the two words also increases. In addition, semantic distance can predict reaction time in semantic categorization tasks and analogy tasks (Rips et al., 1973). For example, the more increased semantic distance between words, the longer it takes to categorize the words as belonging to a

particular semantic relation (e.g., opposites, part-whole) and the longer it takes to complete analogy tasks between sets of semantic relations (e.g., identifying that hand:finger is analogous to foot:toe the same way as it is analogous to tree:branch).

Semantic distance can be easily interpreted when considering the structure of semantic memory. For instance, Quillian (1967, 1969) proposed a model of semantic memory in which every word includes a stored configuration of pointers to other words in memory that represent the word's meaning. For instance, a canary might be stored as a yellow bird that can sing, with "yellow" and "can sing" representing properties of the canary and "bird" representing the broader category to which the canary belongs.

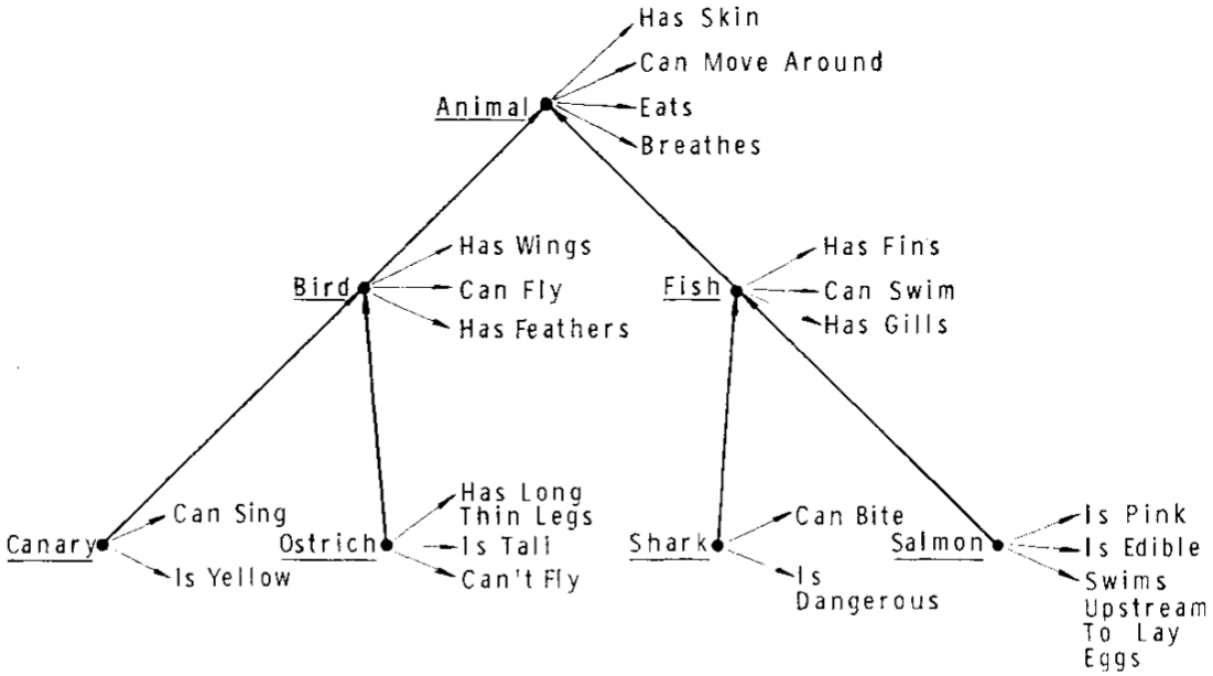


Figure 1. Hypothetical memory structure for a three-level hierarchy (Collins & Quillian, 1969).

As illustrated in Figure 1, the distance in memory structure between a subordinate item ("canary") and a superordinate item ("animal") manifests as a longer reaction time when trying

to reason about the semantic relation between the two items or, more profoundly, when attempting to solve an analogy between the semantic relation that these two items share with an entirely different set of items (e.g., dining room chair:furniture).

When considering the acquisition of semantic relations, research often fails to include the premise of semantic distance as a factor that may contribute to children's ability to reason about particular semantic relations or solve analogy problems containing semantic relations. Collins and Quillian (1969) proposed that the hypothetical memory structure implies a difference between adjectives and nouns when reasoning about the semantic relations between subordinate and superordinate items. It could be that when comparing nouns, children are inherently forced to reason about all the associated properties. In contrast, when comparing adjectives, children only have a single dimension in which to compare them.

Most studies on antonym relation learning have focused on adjective pairs (e.g., *big : small*); however, nouns dominate children's early lexicons compared to verbs and adjectives (Gentner, 1978; Nelson, 1973; Sandhofer & Smith, 2007; Phillips & Pexman, 2015). These findings raise the possibility of similar variability in how children are able to reason about antonyms based on different parts of speech. For example, perhaps children may show earlier success with noun pairs instantiating antonym relation (e.g., *king : queen*). However, though nouns are learned earlier than adjectives, nouns are semantically richer because they hold multiple meanings and share more than one relation with other words, which could make it more difficult for young children to evaluate nouns as compared to adjectives. For example, to generate the opposite of "short" ("tall"), one evaluates the concepts on a single dimension of

length (height); however, to generate the opposite of "king," one could produce "queen" if evaluating based on gender, or "peasant" if evaluating based on economic status.

Therefore, reasoning about antonyms across various parts of speech may follow a different developmental pattern than the acquisition of individual words. Studies assessing the performance of computational models of verbal analogy (e.g., Mikolov et al., 2013; Lu et al., 2019) have compared different semantic relations but not performance across different lexical classes within a single semantic relation of interest. Accordingly, one of the goals of the current studies is to examine human and model performance across different parts of speech: adjectives, nouns, and verbs.

Context as a facilitator for relational reasoning

Semantic relations can be considered abstract categories, for they require children to extract the relation that two words share and generalize it to novel pairs of words that share the same abstract relation. As such, the factors that help children learn relational categories might also facilitate learning semantic relations. Some of the factors that have been found to support the acquisition of categories are relational language, structural alignment, and comparison (Ankowski et al., 2013; Waxman & Markow, 1995; Namy & Gentner, 2002; Gentner et al., 2011; Gentner, 2005).

Studies on relational nouns show that the right syntactic support might clarify their relational nature. For example, a syntactic frame, such as "this is the home of a bird" or "the brother of Y" helps make the argument structure clear (Asmuth & Gentner, 2005). Indeed, research shows that children make more relational responses when they receive a relational

noun label in a relational construction (e.g., "the knife is the dax for the watermelon") vs. when they receive the label in a simple category sentence (e.g., "this knife is a dax") or no label at all (Gentner et al., 2011). Given the relational nature of antonyms or synonyms, it could be that the same kind of syntactic support might benefit children's learning of semantic relations. For example, providing a label for the relation (e.g., "antonym" or a novel word, such as "dax"), or perhaps just enough syntactic support to draw attention to the relational nature of the word, could help children learn the relation more easily.

Another factor that has been shown to facilitate category learning is the use of comparison and contrast, both of which have been shown to provide learners with opportunities to either view various examples of a target category or to compare the target category with non-members of the category (Gentner & Namy, 1999; Ankowski et al., 2013). There is considerable evidence from early naming studies suggesting that children categorize objects based on perceptual features, such as shape or color, rather than on conceptual knowledge (Gentner, 1978; Bowerman, 1976; Smith et al., 1992). Gentner (1978) found that young children extended novel words to objects based on perceptual appearance rather than function, even when the function of the objects was made salient. When preschool-aged children are taught a novel word for a familiar object (e.g., egg) and asked to extend it to new objects, they tend to choose objects that are perceptually similar yet unrelated to the target (e.g., a football) over objects that are conceptually related (e.g., nest) (Baldwin, 1992). In fact, even when 3-5-year-olds were given a choice between an object that shared both perceptual features and category membership and an object that shared only perceptual features with the target object, they were equally likely to select either object (Imai et al., 1994). Despite young children's

tendency to categorize based on shared perceptual features, comparison has been shown to have positive effects on both adult and child learning (Gick & Holyoak, 1983; Gentner & Markman, 1995; Holyoak & Thagard, 1989; Medin et al., 1993).

Studies have shown that comparison promotes categorization, particularly when the category is not bound by salient perceptual commonalities but rather by higher-order commonalities. For example, young children are able to recognize higher order commonalities, such as symmetry, if presented with examples that share perceptual commonalities (Kotovsky & Gentner, 1996). Comparing similar category members is particularly necessary when learning a new relational category, as aligning perceptual features among exemplars allows children to focus on higher order commonalities that would otherwise be less salient (Gentner & Markman, 1994).

According to the structure-mapping theory (Gentner, 1983, 2010), comparison is especially effective at emphasizing relational information, as the structural alignment process helps to highlight the common relational structure that should be attended to rather than features that are not relevant. Though it would seem like learning abstract relations would be best fostered by examining pairs that are different, progressive alignment proves to be effective because the relations are represented in an implicit and context-specific manner (Gentner, 2003; Kotovsky & Gentner, 1996). Specifically, progressive alignment involves starting with concrete, close comparisons and gradually moving to abstract and purely relational comparisons. For example, young children might find it difficult to understand that the relation between a knife and a melon is the same as between an axe and a tree. However, a close similarity pair such as a knife to a melon and a knife to an orange might highlight the similarity between the

corresponding entities and thus highlight the alignment, which would ultimately make the common relation more salient and more easily understood in future examples (Gentner & Medina, 1998; Kotovsky & Gentner, 1996).

Because relational language and structural alignment promote the abstraction of a relation from exemplars, there is reason to believe that the same factors would benefit learning semantic relations, given that semantic relations are inherently abstract. However, the literature on semantic relation learning is still fairly limited in including the factors shown to support relational category and noun learning. For example, when teaching children the "opposite" relation, Phillips & Pexman (2015) minimized the visual similarity (e.g., color, posture, size) of the images that were used in the task to prevent the possibility of children using perceptual similarity as a guide for forming the relation. However, structural alignment would suggest that the opposite approach might be more effective – namely, using images that depict stimuli that share perceptual similarities so that children can focus on the higher order semantic relation that the images share (Gentner, 2003; Kotovsky & Gentner, 1996). Similarly, providing the right syntactic support, including labels and context that emphasizes that pairs of words share a relation, should help children focus on the higher order commonality that pairs of words share and thus facilitate their ability to solve analogy problems involving semantic relations.

Computational Models

While we know that humans learn new categories and relations based on just a few examples, computational models tend to require big data to acquire the ability to complete the same task. However, newer models, such as BART, seek to integrate big data with supervised

learning from small data. Namely, they create semantic features for individual words and then use them to create vectors representing the relations between pairs of words, such as antonyms or synonyms. As such, models like BART are explicitly designed to be able to solve analogy problems involving various types of semantic relations, including the ones discussed in the present studies.

It is important to note the abilities and limitations of these models to inform research that could be tested behaviorally early in development. For example, Lu, Wu, and Holyoak (2019) found that BART and Word2Vec perform less accurately than adults on antonyms despite performing comparably (or even better) on analogy problems involving a large set of semantic relations. As a result, it is essential to examine whether children are also learning semantic relations in similar patterns or, alternatively, what is helping children perform well on analogy problems involving the same relations. Though children eventually do practice analogy problems in schools, there is little research examining how early they become able to complete these problems and what factors are helping them do so. If language input, such as labels, helps children solve these analogies, the same input may also help models.

For example, the model Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), as well as the large language model GPT, are able to receive input such as "is related to" or "is opposite to," which act similarly to the input that children receive and could potentially facilitate their performance on a verbal analogy task. As such, it is important to use the data provided by the models to inform research questions and methods used in behavioral studies and to use the behavioral data to inform ways in which the models can be improved.

Overview of Studies

The proposed studies have a central goal of identifying the factors that contribute to how young children begin reasoning about abstract semantic relations and use them to engage in analogical reasoning. Chapter I aims to elucidate whether and how children generate the antonym pair when given a query word and how this might vary across different parts of speech. Previous research shows that nouns dominate children's early vocabularies compared to adjectives and verbs (Nelson, 1973; Sandhofer & Smith, 2007). Therefore, there is reason to expect differences in antonym generation across lexical classes. Chapter I also implements two vector-based computational models (BART-Gen and Word2Vec), as well as the large language model GPT3.5 in solving the same task.

Chapter II aims to elucidate how much relational labels (i.e., the label "opposites") help children solve analogies involving antonyms in a pictorial analogy task. This experiment separated trials by lexical class to examine potential differences in analogy problems involving different parts of speech. Chapter II implemented BART and Word2Vec, as well as an NLP model BERT, to examine their performance on the same task, as well as how input, such as labels given to children, affects BERT's performance.

Lastly, Chapter III serves as an extension of the experiment in Chapter II by examining whether differences between different lexical classes exist once children are provided with comparison within a trial. Previous work shows that when semantic relations are used to reason by analogy, participants find it easier to think about pairs of words that are more closely related to each other than others. For example, it is easier to see how "a finger is to a hand" and what "a

toe is to a foot" than what "wing is to a bird." That is because the pairs of words finger:hand are more often associated with toe:foot. Computationally, these two pairs of words would be closer in the semantic space than wing:bird. Pairs of words belonging to different lexical classes would naturally be farther away in the semantic space than those belonging to the same class. Thus, Chapter III varies parts of speech within one analogy problem, examining the role semantic distance plays in reasoning about these relations. Based on previous research on semantic distance, I hypothesized that semantically distant pairs of words would be more challenging for children to reason about than semantically near pairs. The same task was also administered to the models in order to examine whether the semantic distance of the mixed analogies proves more difficult for models such as BART or GPT3.5.

Solving classic analogy problems involving semantic relations is certainly an ability required of students once they enter the education system (Common Core State Standards Initiative, 2017). Therefore, it is necessary to examine whether the factors that help children solve spatial or visual analogy problems might also facilitate their ability to solve semantic relations analogy tasks to determine how to help children successfully reason about analogies involving semantic relations that they have not yet been formally taught. For example, research thus far has yet to examine how early children can solve verbal analogies involving semantic relations. Given the evidence showing that children begin to get a sense of these abstract relations around four years of age, the present studies examined whether children in the same age range also used these relations to reason by analogy. This was done through verbal analogy tasks and supporting pictures, which might help children reason about these abstract relations by viewing concrete depictions of the words.

Moreover, analogical reasoning could serve as a tool in semantic relation acquisition. The opportunity to detect the same semantic relation across semantically different examples might help children understand the abstract relation more clearly.

Further, comparing the behavioral studies to computational results elucidates differences in performance on specific analogies where lexical classes are either isolated or not, language support is either given or not, and verbal analogies are either provided or generated. These differences are important, particularly for models like BART and BART-Gen, which are built to mimic relation learning from non-relational inputs (i.e., using embeddings for individual words to create a representation of the relation they share). Elucidating the challenges children experience when first learning how to reason by analogy, as well as the factors that facilitate their ability to do so, will be informative in improving computational models in the future.

Chapter I: The Emergence of Semantic Relation Understanding

Introduction

This chapter focuses on how children and models (BART-Gen and GPT3.5) generate the opposite item in an antonym pair. Experiment IA is the behavioral portion of the first study, and it examines 4- and 5-year-old children's ability to generate antonyms when provided with a query word (e.g., "What's the opposite of big"). Experiment IB is the computational portion of the study and focuses on models such as GPT-3 and BART-Gen. Because this study requires children to generate the antonym pairs, both models will serve as a direct comparison for the behavioral data on the generative task.

The purpose of this study was to examine the accuracy of the responses given by children and models, as well as the patterns of responses. We were interested in the semantic similarity between the default "correct" response and the actual responses generated by both children and models. In order to assess differences across parts of speech, the antonyms used in this task were presented from three lexical classes: adjectives, verbs, and nouns.

IA. Behavioral

Methods

Participants

The study consisted of 37 participants, including 24 4-year-olds (M= 4.458, SD= 0.314) and 13 5-year-old (M= 5.433, SD= 0.404) children recruited through UCLA's Language and Cognitive Developmental Lab. As required by the UCLA IRB, only children whose parents provided consent participated.

Measures

Parents completed a language survey. The language survey included all the words used in the study (including "opposite") to determine whether children have prior knowledge of the words used in the study and whether word knowledge is related to their performance on the analogy task. In addition, parents were asked to complete a demographic questionnaire.

Stimuli and Procedure

This study consisted of 3 training trials and 21 test trials. The three training trials consisted of one pair from each part of speech used (noun, verb, adjective), and the test trials included 7 trials from each part of speech. The word pairs were chosen to ensure that the words

represented valid antonym pairs and were words known by children of this age range. First, the experimenters selected pairs of antonyms from educational booklets that teach antonyms. Then, each word in the pairs was plotted against the proportion of children who tend to produce it by 30 months of age using Stanford's Child Vocabulary Wordbank (wordbank.stanford.edu), which archives data from the MacArthur-Bates Communicative Development Inventory and thus reflects a large dataset of children's English vocabulary acquisition. Only words that 80% or more of children seemed to know by 30 months were used. Although Wordbank tracks language acquisition until 30 months, the youngest children in the present study were 48 months of age. Thus, we took a conservative approach to selecting words children would understand. Then, the pairs were piloted on adult participants who were given one word and asked to generate the other for each pair. Only pairs with high reliability (90%+) were kept in the study.

In the present experiment, words in each pair were labeled as query word 1 or 2 to randomize which words participants were given, with each participant receiving words from either query word list 1 or 2. First, participants completed a warm-up activity to help them feel comfortable responding to the experimenter. During this activity, participants were shown pictures of balloons and asked to name the color of each balloon. Afterward, participants were presented with the first training trial in which the experimenter asked, "What is the opposite of X?" Depending on the response, experimenters confirmed that the participant's response was correct ("That's right! The opposite of X is Y") or corrected the participant's response ("Actually, the opposite of X is Y"). The test trials followed the same format as the training trials, except that participants were not given any feedback after their responses.

Results

We first coded the responses for accuracy strictly, using only the words from the antonym pairs we had initially chosen as the correct response. In subsequent analyses, we consider the synonyms of those words.

We examined whether age affected overall performance and children's performance for antonyms from each part of speech. We found that there was no significant difference in overall performance between four-year-olds (M= .500, SD= .169) and five-year-olds (M= .567, SD=.168), $t(35)$= -0.973, p = 0.341. Similarly, there was a statistically non-significant difference in performance on adjective trials between four-year-olds (M= .657, SD=.206) and five-year-olds (M=.738. SD=147), t (35)= -1.367, p = 0.182; no statistically significant difference in noun trial performance between four-year-olds (M= .571, SD=.218) and five-year-olds (M= .655, SD=.247), $t(35)$= -0.996, p = 0.331; and no statistically significant difference in verb trial performance between four-year-olds (M= .303, SD= .227) and five-year-olds (M= .310, SD= .218), $t(35)$ = -0.086, p = 0.932 (Figure 2).

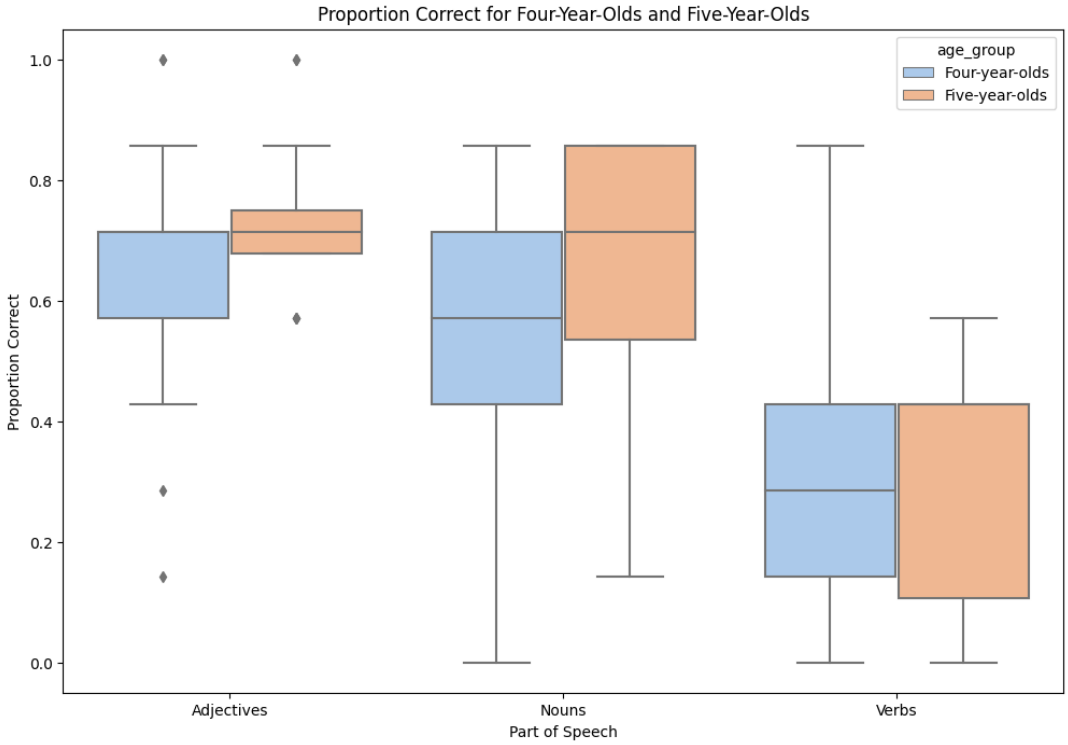


Figure 2. Accuracy on the generative antonym task for both age groups, broken down by part of speech. Each box represents the interquartile range (IQR) of the data, with the median shown as a line inside the box. The whiskers extend to the most extreme data points.

Regardless, we split up the data by age to examine differences in performance between each part of speech. A paired t-test showed no statistically significant difference between adjective (M=.657, SD=.206) and noun (M=.571, SD=.218) accuracy among four-year-olds, $t(25) = 1.964$, $p = 0.061$. However, four-year-olds had a significantly higher performance on noun trials (M=.571, SD=.214) than verb trials (M= .302, SD= .227), $t(25) = 5.260$, $p < 0.001$. Similarly, they performed significantly better on adjective trials (M=.657, SD=.206) as opposed to verb (M= .302, SD= .227) trials $t(25) = 6.699$, $p < 0.001$. These results indicate that four-year-olds performed comparably on trials involving adjectives and nouns but performed significantly worse on verb trials (Figure 3).

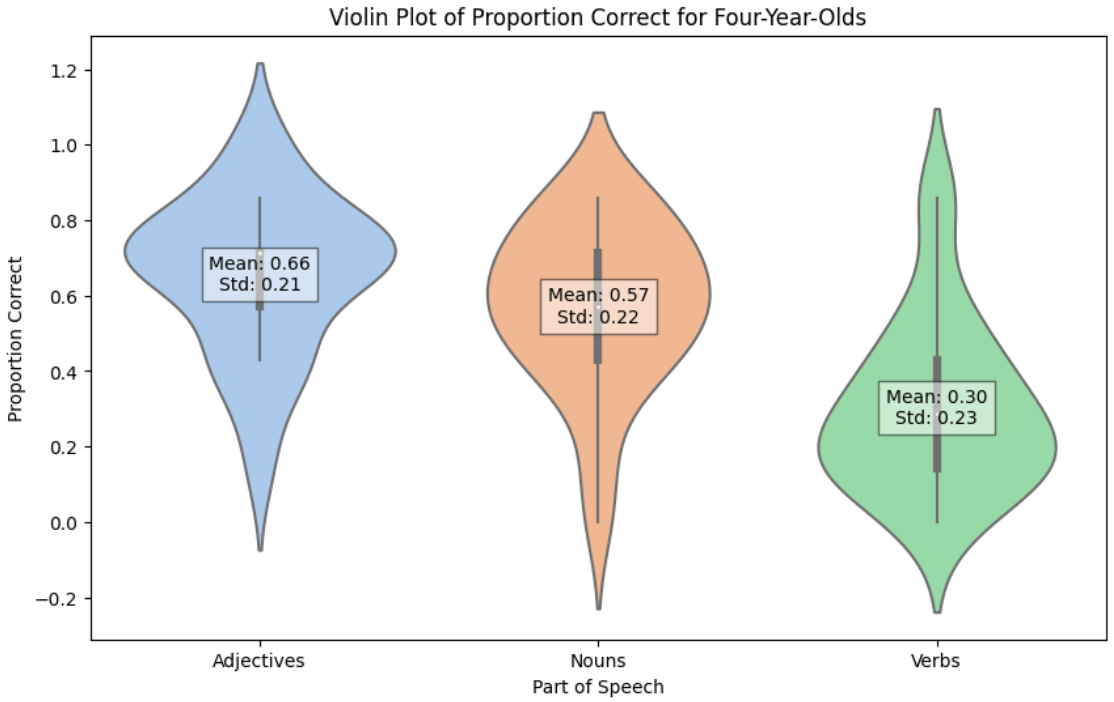


Figure 3. Four-year-old children's accuracy on the generative antonym task, broken down by part of speech. The violin plots reflect the kernal density estimate (KDE), and the means and standard deviations are shown for each part of speech.

We also examined five-year-old children's performance for each part of speech. A paired t-test revealed no statistically significant difference between adjective (*M*=.738, *SD*=.147) and noun (*M*=.655, *SD*=.247) performance among five-year-olds $t(11) = 1.205$, $p = 0.253$. However, similarly to four-year-olds, five-year-olds performed significantly better on noun trials (*M*=.655, *SD*=.247) as opposed to verb trials (*M*= .310, *SD*= .218), $t(11)= 4.994$, $p < 0.001$. They also performed significantly higher on adjective trials (*M*=.738, *SD*=.147) than on verb trials (*M*= .310, *SD*= .218), $t(11) = 9.95$, $p < 0.001$. These results indicate that among five-year-olds, there was no significant difference between adjective and noun accuracy, but they were able to generate adjective and noun opposites significantly better than they could verb opposites (Figure 4).

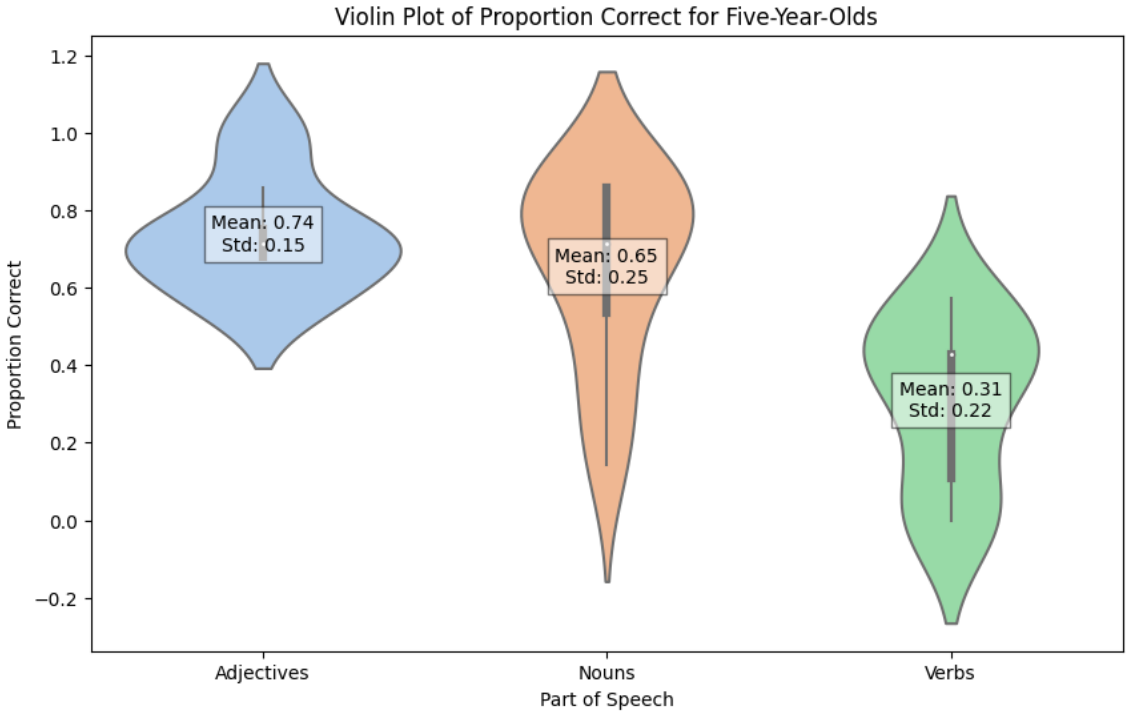


Figure 4. Five-year-old children's accuracy on the generative antonym task, broken down by part of speech. The violin plots reflect the kernal density estimate (KDE), and the means and standard deviations are shown for each part of speech.

When examining correlations between parent reports of children's language knowledge involving the words specifically used in our experiment, we found that overall performance on the antonym generation task and the knowledge of the words we expected children to know in order to complete this task was significantly correlated $r(35)= 0.41$, $p=0.01$ (Figures 5 and 6). However, there was no significant correlation between overall performance on the task and whether the parents reported if the children knew the word "opposite" $r(35)= 0.19$, $p=0.26$ (Figure 5).

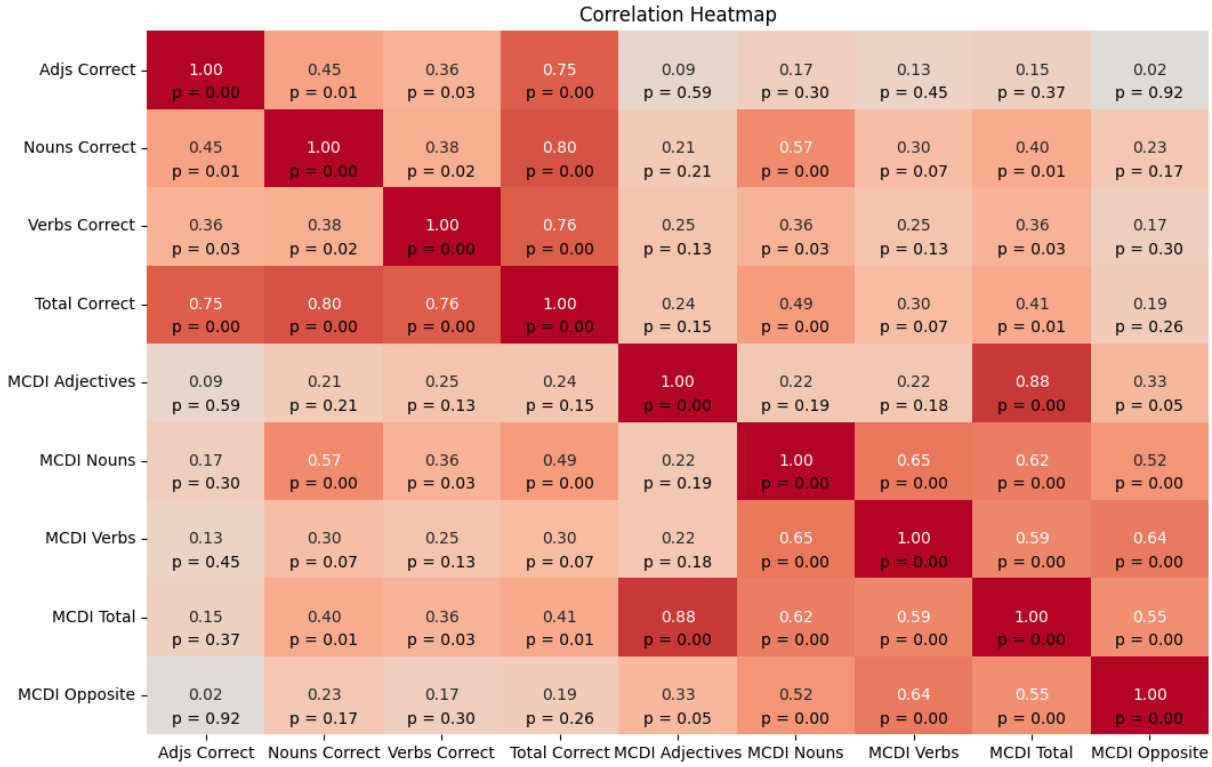


Figure 5. Correlations between accuracy on the generative antonym task and parent reports of children's word knowledge. The accuracy on the task is shown as a total, as well as broken down by part of speech. Parent reported language ability is reflected by the MCDI label. "MCDI Opposite" reflects parents' reports of whether children know the word "opposite."

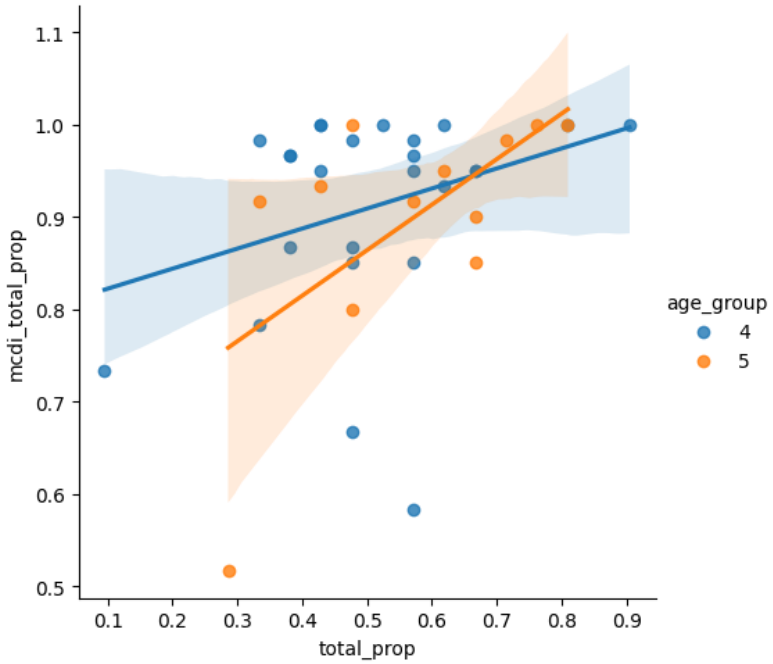


Figure 6. Correlations between accuracy on the generative antonym task, split by age group, and parent reports of children's knowledge of the words used in the task.

Recoded Results

We recoded all the responses to account for responses that do not directly match our original pairs but are synonyms of those words or other reasonable potential antonyms ("legal" responses). Similar to the original coding, we found no significant difference in the adjective proportion correct between four-year-olds (M=.709, SD=.196) and five-year-olds (M= .750, SD= .151), $t(35)$= -.708, p = .485; no significant difference in the noun proportion correct between four-year-olds (M= .600, SD= .222) and five-year-olds (M=.690, SD=.250), $t(35)$= -1.068, p =.299; no significant difference in the verb proportion correct between four-year-olds (M= .434, SD=.239) and five-year-olds (M= .440, SD=269), $t(35)$= -.068, p=.946; and no significant difference in the total proportion correct between four-year-olds (M= .581, SD= .168) and five-year-olds (M= .627, SD= .188), $t(35)$= -.720, p=.480. These findings suggest that even with lenient coding, the discrepancy between four- and five-year-olds' performance on the antonym generative task remains non-significant (Figure 7).

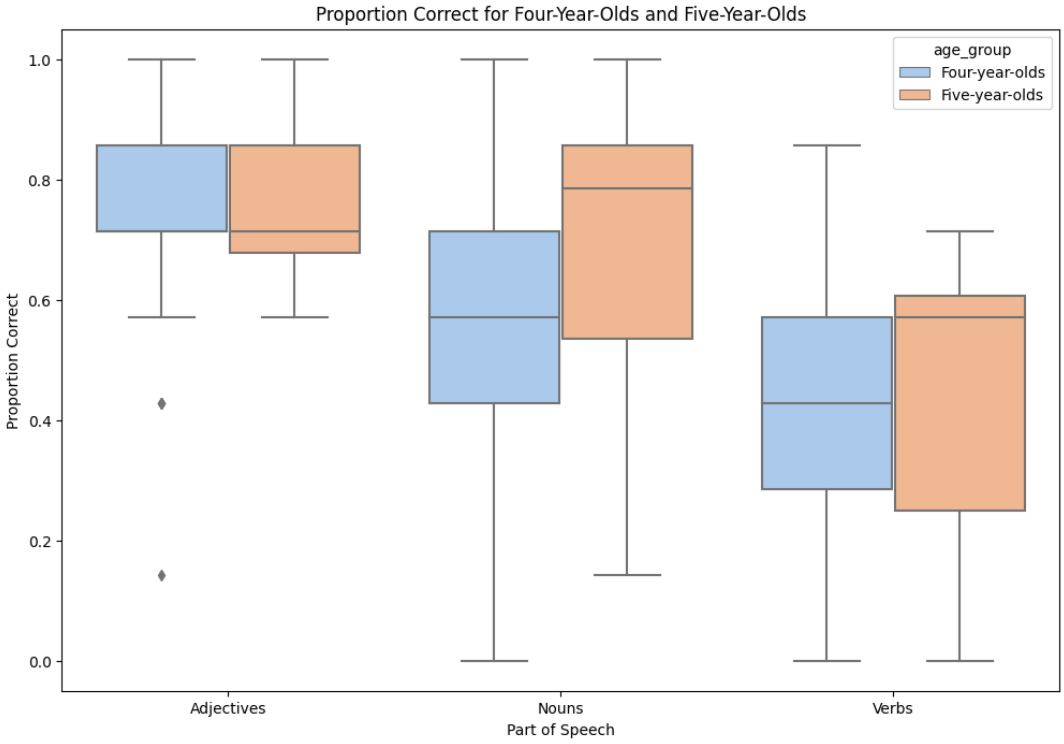


Figure 7. Accuracy on the recoded generative antonym task for both age groups, broken down by part of speech. Each box represents the interquartile range (IQR) of the data, with the median shown as a line inside the box. The whiskers extend to the most extreme data points.

When testing strictly for four-year-olds' performance, a paired t-test showed that performance on adjective trials was significantly higher (M= .709, SD= .196) than on noun trials (M=.600, SD= .222) among four-year-olds, $t(24) = 2.671$, $p = 0.013$. Additionally, four-year-olds performed significantly better on nouns (M=.600, SD= .222) than on verbs (M=.434, SD= .239), $t(24) = 3.112$, $p = 0.005$. They also performed better on adjectives (M= .709, SD= .196) than on verbs (M=.434, SD= .239), $t(24) = 5.331$, $p < 0.001$. These findings suggest that even with more lenient coding, four-year-olds performed significantly better on nouns and adjectives compared to verbs (Figure 8).

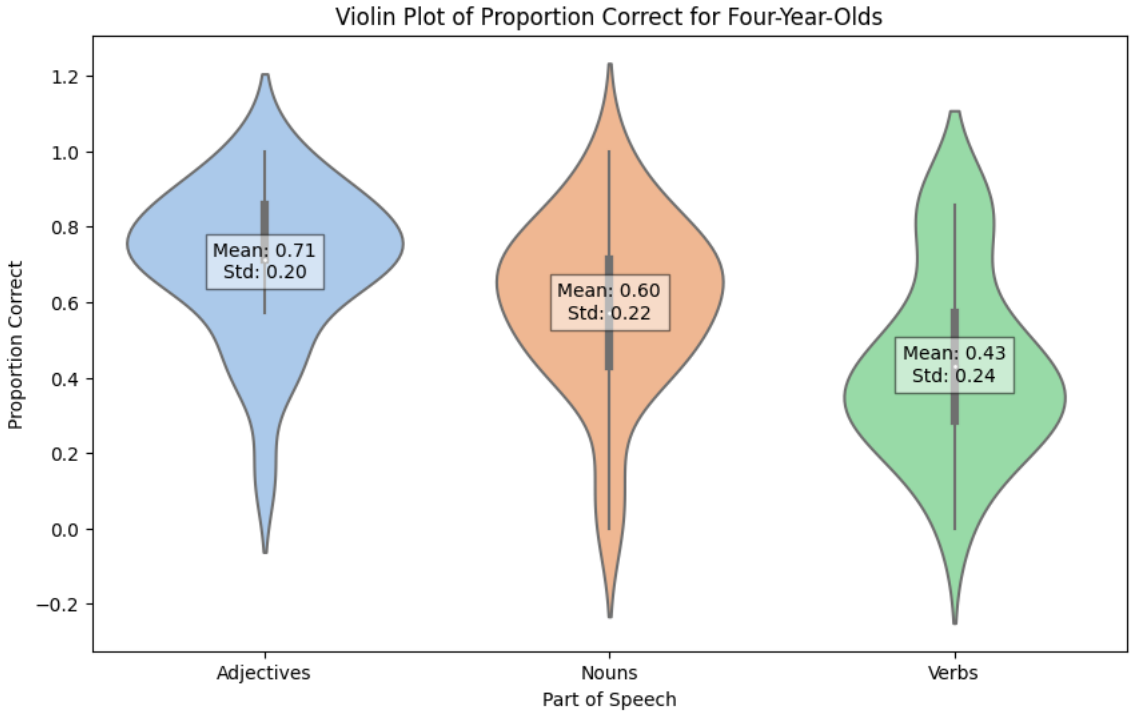


Figure 8. Four-year-old children's accuracy on the recoded generative antonym task, broken down by part of speech. The violin plots reflect the kernal density estimate (KDE), and the means and standard deviations are shown for each part of speech.

When examining the performance of five-year-olds, we found no statistically significant difference between adjectives (M= .75, SD= .151) and nouns (M= .69, SD= .25), $t(11) = 0.861$, $p = 0.408$. However, similarly to the original coding, the proportion of correct responses on nouns (M= .69, SD= .25) was significantly higher than that on verbs (M= .44, SD= .269), $t(11) = 3.540$, $p = 0.005$. Additionally, performance on adjectives (M= .75, SD= .151) was also significantly higher than that on verbs (M= .44, SD= .269) among five-year-olds, $t(11) = 5.613$, $p < 0.001$ (Figure 9). These findings suggest that even when coding for responses that did not fit the originally chosen pairs and instead included synonyms or other valid antonyms, there was no significant difference between the proportion of correct responses for both adjectives and nouns among five-year-olds, but they performed significantly better on nouns and adjectives as opposed to verbs.

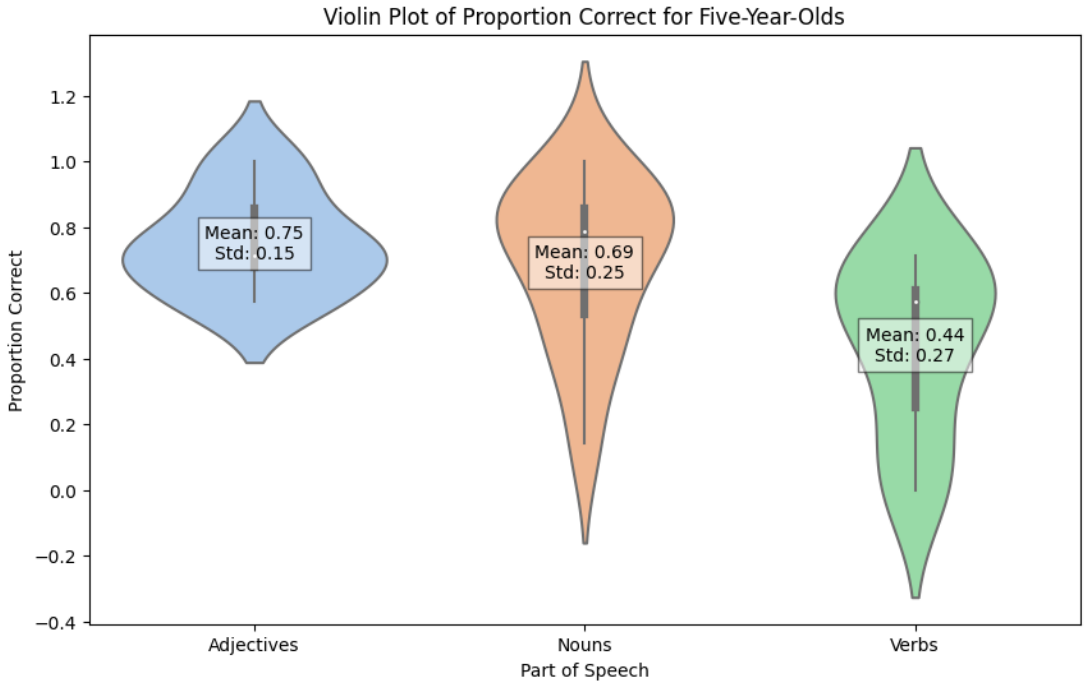


Figure 9. Five-year-old children's accuracy on the recoded generative antonym task, broken down by part of speech. The violin plots reflect the kernal density estimate (KDE), and the means and standard deviations are shown for each part of speech.

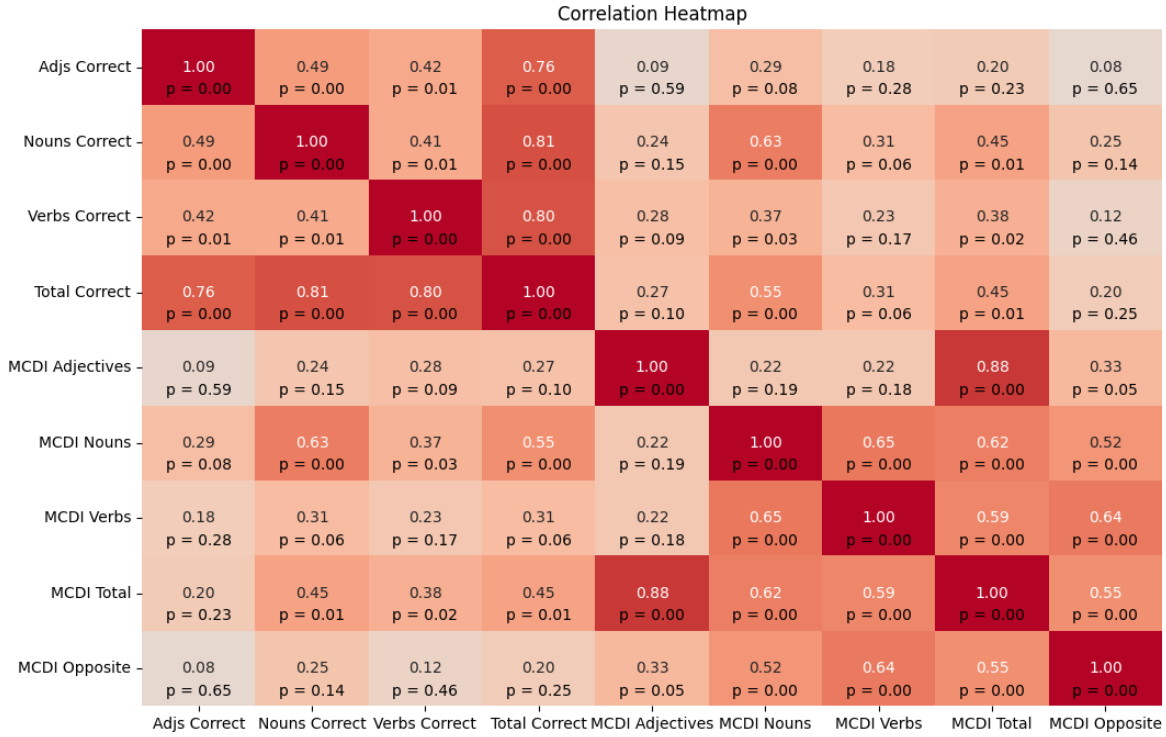


Figure 10. Correlations between accuracy on the recoded generative antonym task and parent reports of children's word knowledge. The accuracy on the task is shown as a total, as well as broken down by part of speech. Parent reported language ability is reflected by the MCDI label. "MCDI Opposite" reflects parents' reports of whether children know the word "opposite."

We computed Pearson correlation coefficients between parent reports of children's language knowledge involving the words specifically used in our experiment and the words the children generated even when they were recoded for possible correct alternatives, and we found that overall performance on the antonym generation task and the knowledge of the words we expected children to know in order to complete this task was still significantly correlated $r(35)= 45$, $p=0.01$ (Figures 10 and 11). However, there was still no significant correlation between overall performance on the task and whether the parents reported if the children knew the word "opposite" $r(35)= 0.20$, $p=0.25$ (Figure 10).

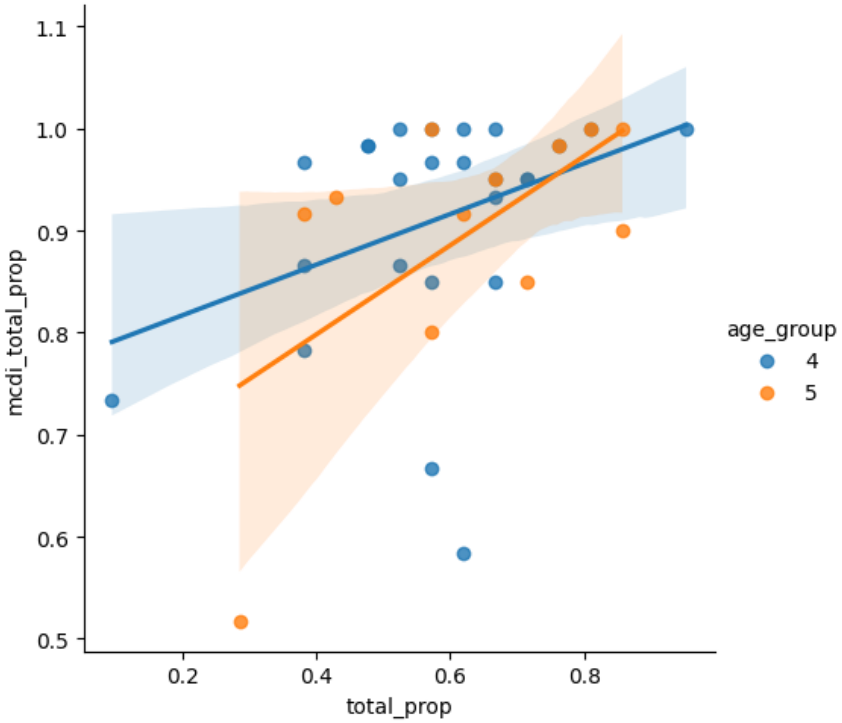


Figure 11. Correlations between accuracy on the recoded generative antonym task, split by age group, and parent reports of children's knowledge of the words used in the task.

Apart from recoding the data to account for target word synonyms or other potentially correct antonyms, we wanted to examine the semantic distance of the words that children

generated relative to the target word. This way, we can computationally measure whether children's responses are relatively semantically close to the target word. To do this, we retrieved the word2vec (w2v) embeddings of each query word, each target word, and each generated response word. Using the Python package t-SNE, we were able to convert the w2v embeddings to 2 dimensions to plot them on a scatterplot. With the two conditions collapsed across participants, we have 14 target adjectives, 14 target nouns, and 14 target verbs. Using these plots, we can observe the distribution of the generated words around the target words and whether clusters form around the target words. In the scatterplots in Figure 12, we observe clear clusters of generated words formed for nouns and adjectives, and a broader distribution for verbs. This would suggest that the words children generate for the adjective and noun antonyms are closer in the semantic space than those they generate for verbs.

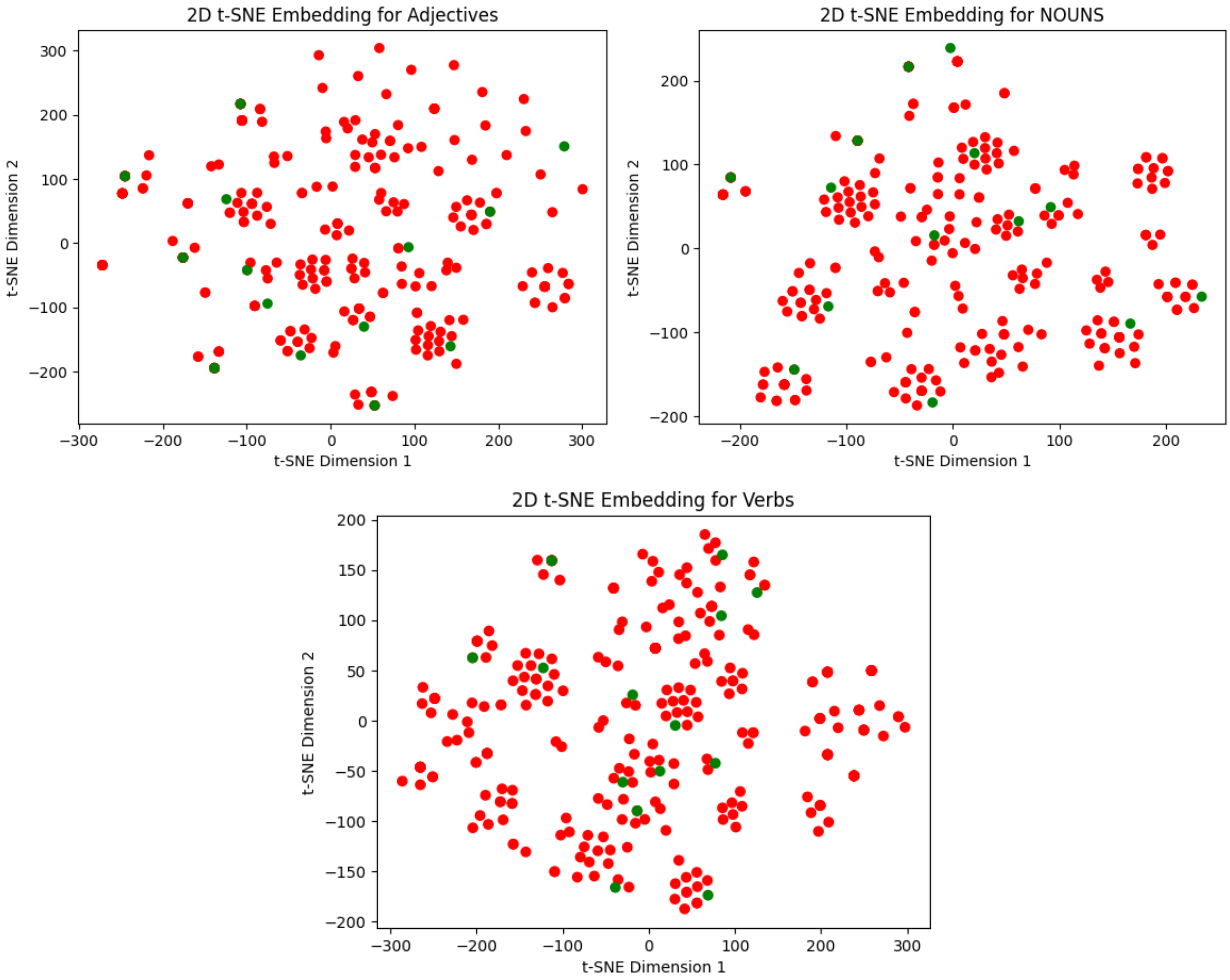


Figure 12. W2V embeddings for the target and generated words in the generative antonym task, collapsed to two-dimensional space. The plots are split by part of speech, and in each plot the green dots depict the 14 possible target words for each part of speech, and the red dots reflect the words that the children generated. The jitter is set to 0 for target words and 1 for the generated words. In some cases, if the generated word is the target word, the dots overlap.

Next, we computed the cosine distance between each unique target word and each corresponding generated word to examine whether the clusters we observe in our TSNE plots do indeed coincide with the average semantic distance between each target word and each corresponding generated word. We grouped the target words by part of speech and found that there is a significantly higher average semantic distance between target words and generated words for verbs (M=.458, SD=.7) as opposed to adjectives (M= .244, SD=.193), $t(26)$= -3.1559, p=

.004, and a significantly higher average semantic distance for verbs as opposed to nouns (*M*=.238, *SD*=.182), *t*(26)= -3.365, *p*= .0024. However, much like the accuracy analyses, we found no significant difference between the average semantic distance between target words and generated words for adjectives (*M*= .244, *SD*=.193) and nouns (*M*=.238, *SD*=.182), *t*(26)= .098, *p*=.923 (Figure 13).

Figures 14-16 depict the frequency and distribution of the words children generated, corresponding to each target word across the three parts of speech.

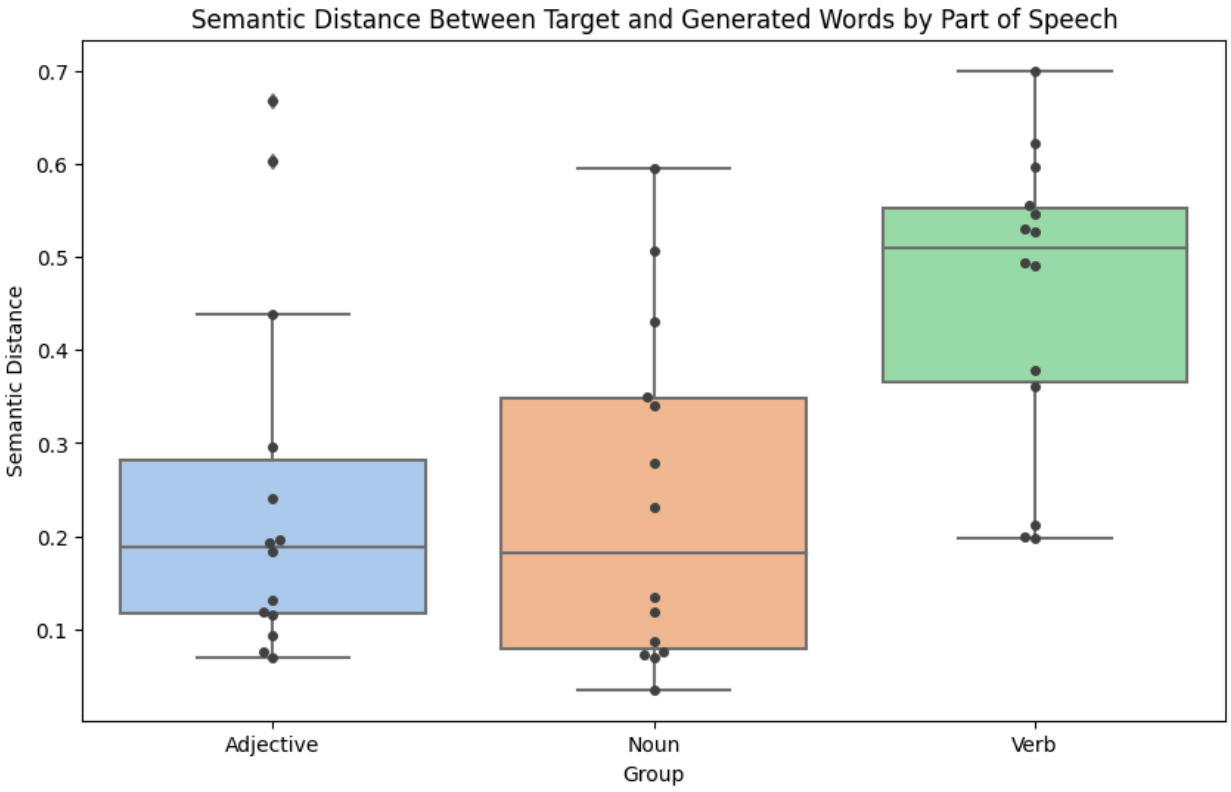


Figure 13. The cosine distance between the target words and the words that children generated on the antonym task, separated by part of speech. Each box represents the interquartile range (IQR) of the data, with the median shown as a line inside the box. The whiskers extend to the most extreme data points.

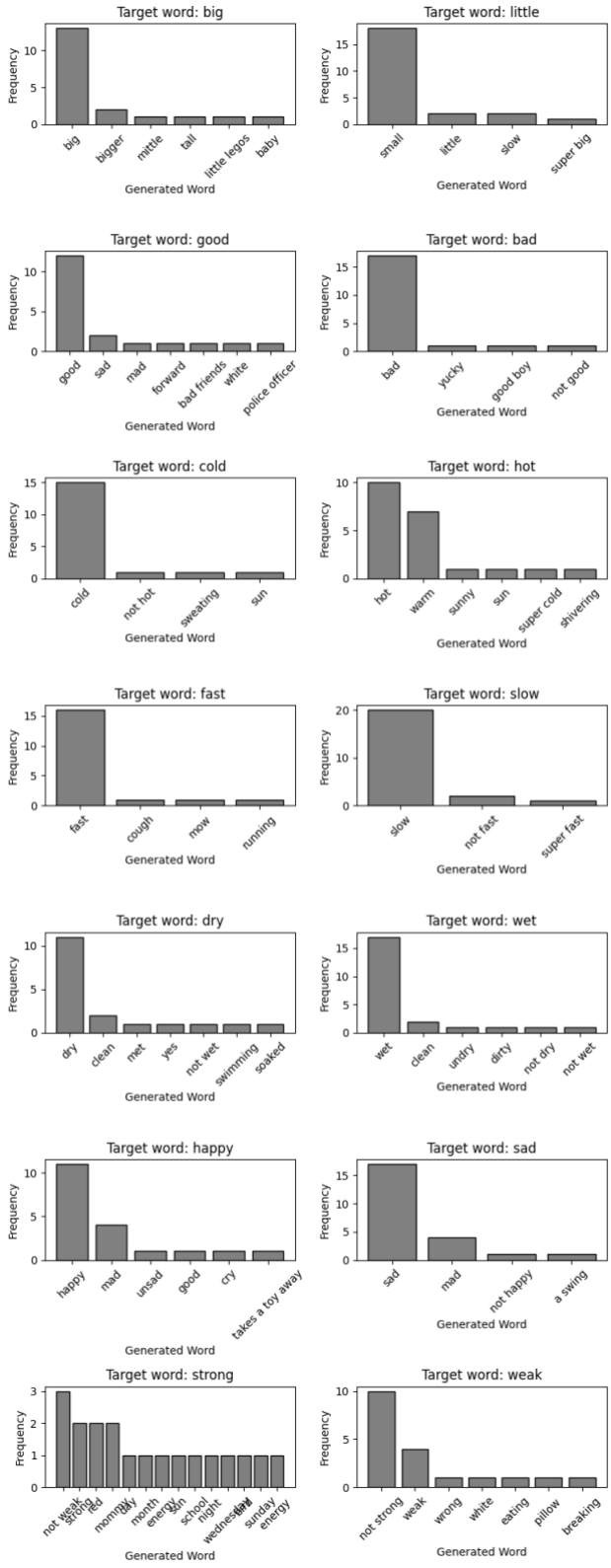


Figure 14. Frequency of each of the words that children generated on the antonym task, according to each adjective target word. The target words on each row correspond to the same relation (e.g., cold:hot, happy:sad). If the target is cold, the source word is the opposite.

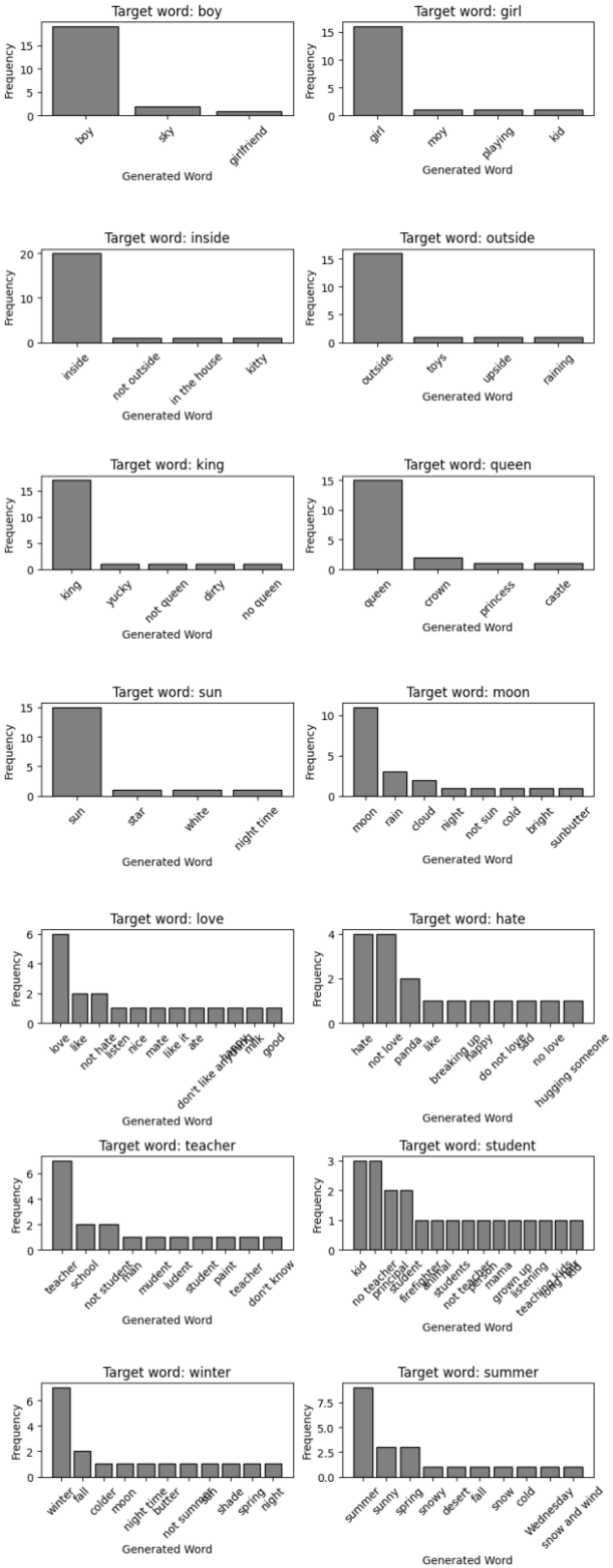


Figure 15. Frequency of each of the words that children generated on the antonym task, according to each noun target word. The target words on each row correspond to the same relation (e.g., winter:summer, love:hate).

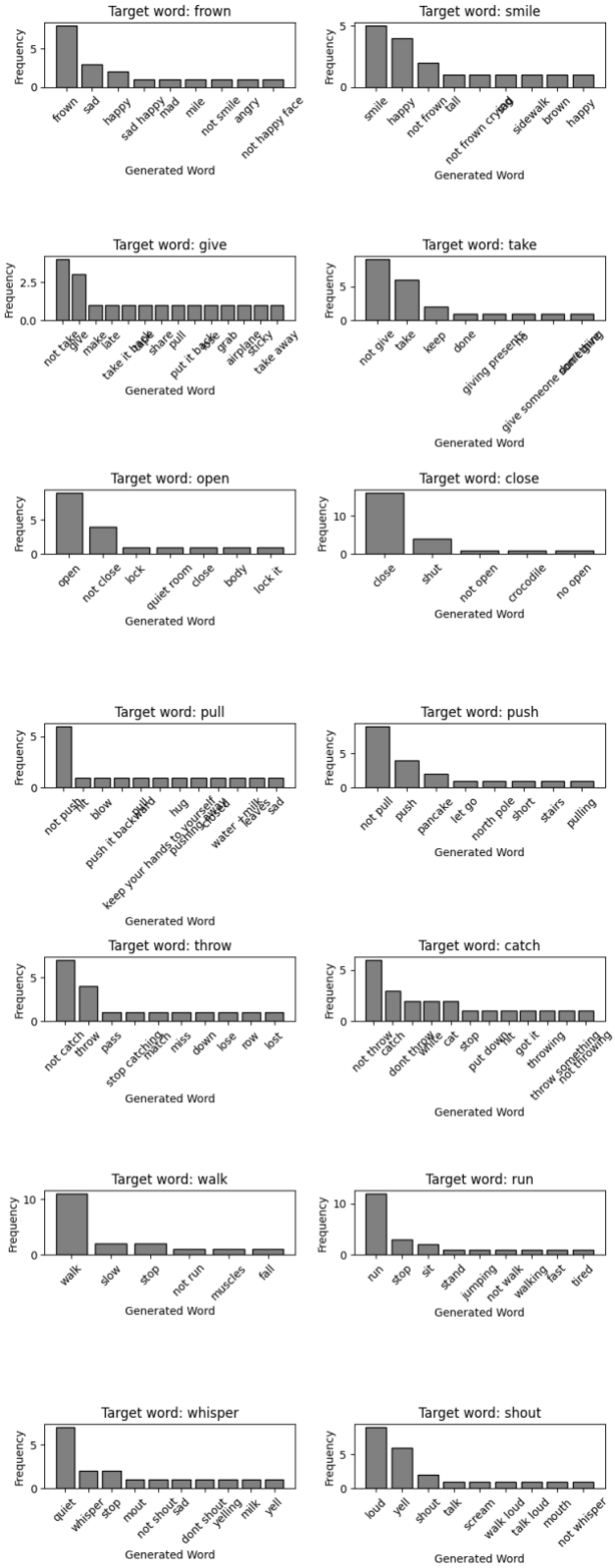


Figure 16. Frequency of each of the words that children generated on the antonym task, according to each verb target word. The target words on each row correspond to the same relation (e.g., pull:push, frown:smile).

IB. Computational

In this portion, we implemented the model BART-Gen, which can receive language input similar to that given to children (e.g., what is opposite of X?") in order to generate the target word in an analogy. BART-Gen is intended to match adult performance in correctly identifying semantic relations.

BART-Gen uses the relational representations generated by BART (*Bayesian Analogy with Relational Transformations*), a model of relation learning that forms representations of relations from vector representations of individual word meanings (Lu et al., 2019). Namely, it relies on a distributed vector representation of the relation between words A and B (R_{AB}), which corresponds to one of the 270 distinct relations learned by BART. After being given the input for word C, the model generates 270 predictions of word embeddings for the target word D, and computes a weighted average of the set of generated embeddings scaled by a normalized relation vector. To successfully generate the word embedding for target word D, the model must make predictions from the relations for which A and B are a positive example, compared to the relations for which A and B are a negative example (Ichien, N., Kan, A., Holyoak, K. J., & Lu, H., 2022). The individual word vectors used by BART to form the relational vectors are provided by Word2Vec (Mikolov et al., 2013). As such, we also included W2V's performance in our findings as a reference point.

We also conducted the same task on GPT3.5 through ChatGPT. GPT3.5 is an advanced artificial intelligence model designed by OpenAI for natural language processing tasks. While OpenAI has not disclosed the number of parameters, it is safe to assume that it is larger than its

predecessor, GPT-3, which comprised 175B parameters. Given its advanced training, it should also serve as a good comparison for children's performance on the antonym generation task.

Results

When provided the word pairs from our study, BART-Gen and W2V each provide a ranking for the target word in relation to the query word, with a ranking of 1 signaling that the target word is the first choice for the models' in terms of what constitutes the opposite of the query word, followed by all the words which they are trained on in the context of our relations of interest. In order to mark what constitutes a correct or incorrect response in our dataset, we used a cut-off limit of 10, marking performance as being correct as long as the models generated words in the top 10 rankings. For GPT3.5, we marked the words as correct or incorrect while allowing alternative responses to be correct if they were logical. This coding scheme was intended to resemble that used with children, in which their responses were recoded for accuracy while allowing for legal words. All models were administered one of the query words and then reassessed while the second query word was administered. In our results, performance is shown averaged across the two query word simulations.

Table 1 shows that performance on the antonym generative task varied across the models, with GPT3.5 showing a perfect performance on each part of speech. BART-Gen and W2v each show the highest performance on nouns, followed by adjectives and verbs at a similar accuracy rate.

Table 1. Proportion model performance on the generative antonym task.

	Models	Adjectives	Nouns	Verbs
Relation model	BART-Gen	0.57	0.86	0.50
NLP models	W2V	0.50	0.86	0.57
	GPT3.5	1	1	1

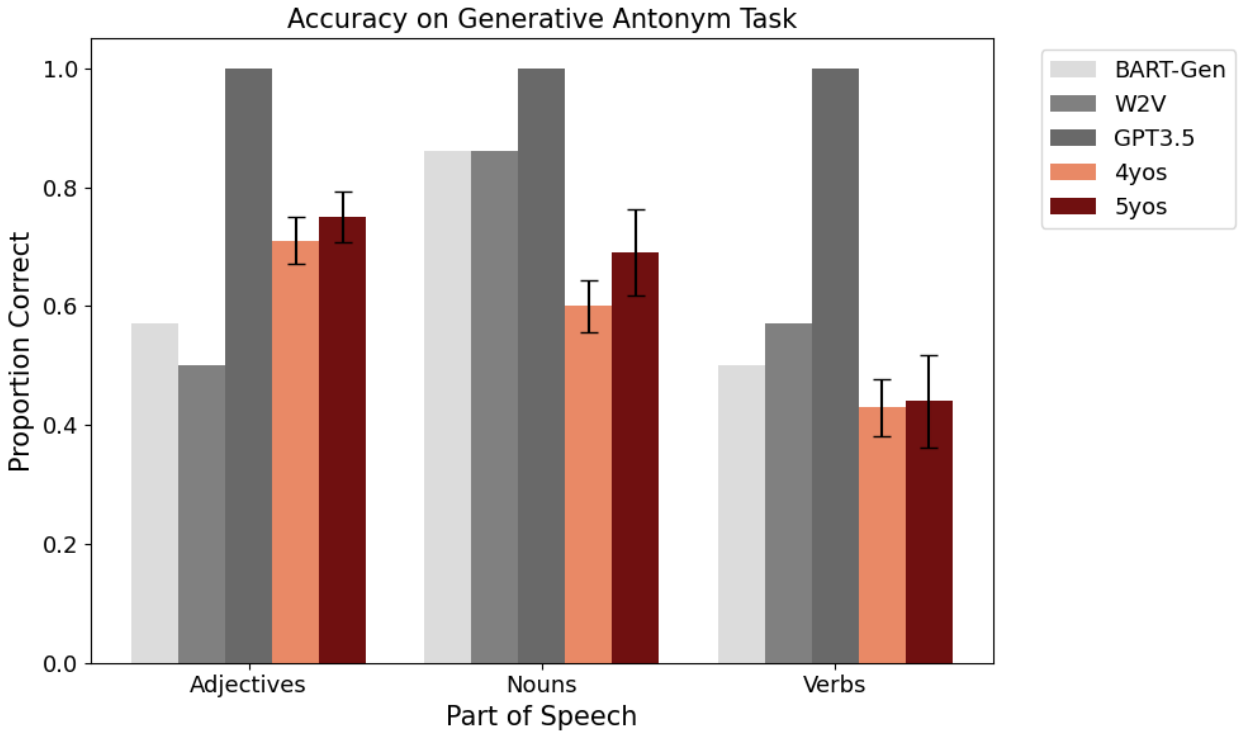


Figure 17. Accuracy on the generative antonym task for both models and children.

Our simulations show that BART-Gen and W2V perform worse than children on adjective pairs, but better on noun and verb pairs. GPT3.5, on the other hand, performs considerably better than children on all parts of speech (Table 1, Figure 17).

Chapter I Discussion

In Chapter I, I reported on both a behavioral and computational approach to examining antonym word-pair generative responses. In Chapter IA, I presented four- and five-year-old children with query words belonging to three parts of speech (adjectives, nouns, and verbs) and prompted them to generate the corresponding antonym. The behavioral task showed that children of both ages demonstrate overall high performance on both adjectives and nouns, and lower performance on verbs. Even when allowing for a multitude of "legal" responses that were not our intended antonym target words but semantically similar and conceptually a good fit for an antonym pair, the same patterns were present for both ages, though performance increased overall for all parts of speech. These findings somewhat match children's lexical development patterns as children's early vocabularies contain more nouns and verbs earlier than adjectives (Nelson, 1973; Sandhofer & Smith, 2007). However, many adjective antonym pairs tend to vary on a single dimension, whereas nouns and verb pairs are typically more semantically complex.

When examining the variability in the generated responses of the children for each part of speech, we found that verbs especially tended to have much more variability than adjectives and nouns. This could be due to the subjective nature of verb antonyms because they can be evaluated on multiple dimensions. When children did not know the correct response for a verb antonym, they tended to simply negate the original query word (e.g., catch:not catch, hate:not hate). This suggests that when children do not know the correct antonym pair, they prioritize a "different" relation in which they know that the target word is something other than the query word, even if they do not know the correct response.

When looking at the word embeddings for each target word and each generated word, we found that particularly for adjectives and nouns we tended to see more clusters being formed around the target words suggesting that even if children are not necessarily generating the right target word, they are generating words that are semantically close to the target words. The verb embedding plots also show some clusters forming around the target words. However, we see much more variability in where each word embedding is in the two-dimensional semantic space as opposed to nouns and adjectives. This further demonstrates the findings from the statistical test suggesting that accuracy tends to be higher for adjectives and nouns and lower for verbs. Overall, these findings suggest that there is variability across parts of speech and that even though children seem to have a solid understanding of the antonym relation and how it applies to various instantiations of the relation, they still have difficulty generating words to complete antonym pairs, especially if they might not yet have high familiarity with them due to their age.

To measure whether familiarity with the words played a role in children's overall performance on the task, we found that familiarity with the words, as reported by parents, is significantly correlated with overall performance on the antonym generation task. Interestingly, however, there was no significant correlation between overall performance on the antonym task and whether the children knew the word opposite as reported by parents. Because overall parents reported that children knew most of the words used in our experiment, this suggests that lexical development is correlated with the ability to reason about relations between the words. When it comes to the word *opposite* specifically, it could be that children either know the word but do not yet completely understand how it applies to different instantiations, or it could

be that parents were either not aware that their child knew the word *opposite* or were not completely accurate when completing our questionnaire, which is a general limitation of parent language reports.

When examining the model simulation results against those of children, we see a different pattern. BART-Gen and W2V perform comparably to, or better, than children on noun and verb pairs, yet considerably worse on adjective pairs.

Chapter II: Analogies Involving Antonyms

Introduction

This chapter focuses on how children and models (Word2Vec, BART, BERT) complete verbal analogy problems using antonyms in a pictorial task. Experiment 2A is the behavioral portion of the final study, which examines how 4- and 5-year-old children perform when given an example pair of antonyms and then shown two more pairs (one antonym, one a distractor) and asked to choose which of the pairs corresponds to the example, thus having to solve an analogy problem successfully. For Experiment 2A, we predicted that children in the relational label condition would perform better than children in the no label condition. Experiment 2B focuses on model performance on the same task.

The purpose of the experiments in this chapter is to expand past the ability to generate antonyms and instead examine how children and models use them to solve analogy problems. Moreover, the experiments are intended to investigate the factors that facilitate this ability, including language cues in the form of labels (i.e., "opposites") and any potential differences across parts of speech. Because children tend to learn nouns first, followed by adjectives and

verbs, we were interested in examining whether such differences affect young children's ability to draw analogies across antonym pairs that involve each part of speech individually.

IIA. Behavioral

Methods

Participants

We tested 107 participants, including 54 four-year-old (M = 4.42, SD = .27) and 53 five-year-old (M = 5.55, SD = .26) children were recruited through the Language and Cognitive Development Lab at the University of California, Los Angeles (UCLA) either through Lookit, an online recruitment platform hosted by MIT, or through the UCLA Developmental Subject Pool. In accordance with the UCLA Institutional Review Board policies, only children whose parents granted formal consent participated. Data collection was completed entirely online using Zoom.

Measures

Parents completed a language survey adapted from the MacArthur Bates Communicative Development Inventory in which they were instructed to identify the words their child produces. This survey included words that appeared in the analogy task (e.g., "happy" and "opposite") to determine whether the children had prior knowledge of the words used in the study and whether their word knowledge was related to their performance on the analogy task.

Materials

The pictorial analogy task followed the format of a Relational Match-to-Sample (RMTS) task (see Figure 1) (see Appendix N for full set of word pairs and Appendix D for full set of

pictures). Children were allowed to simultaneously compare a source pair exemplifying a contrast relation to a target pair also exemplifying a contrast relation but on a different dimension than the source (e.g., size versus cleanliness) and a distractor pair that was semantically unrelated. Thus, the relational match between the source and target was at the abstract level of antonym rather than at the level of a more specific relational contrast.

The pairs corresponding to antonyms and the distractor pairs were pictures of people, familiar animals, and objects. As illustrated in Figure 1, the objects used in the target and distractor were from the same category, which differed from the category used for the source objects. The contrastive relations used in the task could be expressed as either adjectives (e.g., *happy : sad :: dry : wet* or *tired : dry*), nouns (e.g., *friends : enemies :: teacher : mother* or *teacher : student*), or verbs (e.g., *open : close :: build : destroy* or *build : stop*).

Stimulus validation

The antonym word pairs were sourced from educational websites and subsequently verified on WordBank (cite), a database of words produced by at least 50% of children by 30 months of age. Word pairs were chosen by selecting only those known to over 80% of 30-month-olds. To determine the validity of antonym pairs, we conducted a Google Form survey with adults to validate which words are considered "opposites." Two forms were created, each of which included one of the words in each pair. Twenty-five UCLA undergraduates were asked to generate the antonym for each word on a list, and only pairs with reliability over 95% were chosen for the final list.

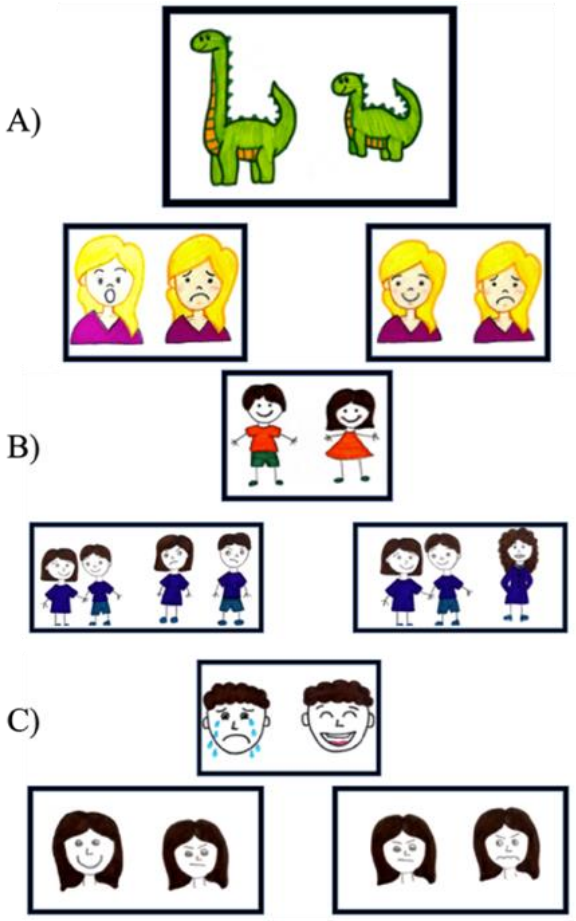


Figure 18. Examples of three trials on the pictorial antonym analogy task, illustrating the three lexical classes used in the task. A: An adjective source pair exemplifying a contrastive relation (big : small), with a distractor pair (surprised : sad) on the left and the correct option (happy : sad) on the right. B: A noun source pair (boy : girl), with the correct option (friends : enemies) on the left and distractor pair (friends : mother) on the right. C: A verb source pair (cry : laugh), with the correct option (smile : frown) on the left and a distractor pair (frown : hate) on the right.

Procedure

Children received three training trials (one per part of speech) and thirty test trials (ten per part of speech), all within subjects. Children were assigned to one of two conditions (between subjects): the Label and No-Label conditions. On each trial in the Label condition, children were told that the animals/objects/humans depicted in the source pair were "opposites" of each other (e.g., "This is dirty, this is clean. Dirty and clean are *opposites"*) in both practice and test trials. The words used to describe objects were either adjectives, nouns, or

verbs. In the No-Label condition, children were not given a label for the abstract relation in any of the trials. Instead, children were only provided with verbal descriptions of the individual objects (e.g., "This is dirty and this is clean"). The pictures used in the study were not intended to be the primary sources of information, but rather, their use was designed to facilitate children's understanding of the key words in the study and to provide a concrete representation of each of them. As such, the verbal descriptions of the individual objects were given for every image that appeared in the study, regardless of the condition.

For each condition, we created five versions of the task to semi-randomize which source pairs were matched with which target/distractor pairs. Because some picture pairs were repeated across trials, combinations were semi-randomized so that a target pair never appeared earlier as a source pair. For example, if a pair based on *big*/*small* was used as the source pair in the first test trial, that pair was never used afterward as a target pair. The part of speech was kept consistent among the source, target, and distractor pairs for every trial. In addition, the display position of the target/distractor pair was randomized between trials such that the correct pair appeared on the left side of the screen for half of the trials and on the right side of the screen for the other half. Children were randomly assigned to one of the five versions within each condition.

Practice trials

To begin, children were shown a source picture showing two animals, objects, or humans depicting a pair of antonyms (e.g., a big balloon and a small balloon; see Figure 1). The experimenter labeled the pictures, emphasizing the words that depicted the contrastive relation (e.g., "This is *big*, this is *small*. Big and small"). Afterward, simultaneously, the experimenter

provided two more images that respectively depicted either a target pair of antonyms (e.g., a *clean* pig and a *dirty* pig) or a distractor pair of semantically unrelated words, one of which was kept consistent with the antonym pair (e.g., a *clean* pig and a *sad* pig). The experimenter always described each of the pictures, emphasizing the key words (e.g., *clean* and *dirty*). The participants were asked, "Which one is like this one (pointing to the source picture)?" Children were given feedback: either told that they were correct or told the correct answer if the child provided an incorrect one.

Test trials

The format of test trials was identical to that of training trials, except that the children were not given any feedback regarding their answers on each trial. The animals/objects/humans shown in the target and distractor pictures were always kept consistent in color and category (animals, objects, or humans). The target and distractor pictures differed in color and category from the source picture. These constraints ensured that children could not simply choose the picture that was most similar to the source picture based on the features of individual objects.

Results

To analyze children's accuracy in selecting the correct target, we implemented a Bayesian logistic regression model using the R package brms (Burkner, 2018). We tested hypotheses by fitting a logistic regression model predicting responses on the analogy task based on the interaction between condition (Reference = No Label) and age (Reference = four-year-olds). This model included group-level effects of subject and item and allowed for heterogeneity in the intercepts of the effects of condition and age. The model also included a grouping of the

item types into three parts of speech to analyze differences between analogies based on nouns, verbs, and adjectives. For the prior distributions in our model, we used a uniform (i.e., uninformative) distribution for the main effects and interaction coefficient, and used a t(3,0,2.5) for the random intercepts and their standard deviation. Specified in brms syntax, the model is:

Response ~ Condition*Age + PartofSpeech + (1| Subject) + (1| Item)

These analyses revealed that being in the older age group and being given the relation label of "opposite" predicted higher accuracy on the analogy task (b= 0.45, 95% CI [-0.39, 1.31]) (see Table 2 and Figure 19). Moreover, the pattern of results suggested that labeling the antonym relation was particularly effective for five-year-old children. In contrast, labeling the antonym relation made less of a difference for four-year-old children overall. Separated by part of speech, results show that being given a label helped four-year-olds do significantly better on verb trials (M= .672, SD=.191) compared to the no-label condition (M=.536, SD= .2196), t(52) = -2.44, p = .018). We found that in the no-label condition, four-year-olds were not able to perform significantly above chance on verb trials, t(52) = .820, p = .42, indicating that this abstract analogy task is too difficult for four-year-olds to solve with above-chance accuracy for verbs when not given a relational cue. Similarly, reliable performance on other versions of RMTS problems is not observed prior to age five (Hochmann et al., 2017).

Table 2. Estimates of Posteriors for Bayesian Logistic Regression Model.

Population-Level Effects	Estimate	Est. Error	Lower 95% CI	Upper 95% CI
Intercept	-1.51	1.37	-4.17	1.15
Condition	-1.40	1.94	-5.33	2.42
Age	0.51	0.30	-0.06	1.09
Part of Speech – Noun	-0.22	0.23	-0.68	0.23
Part of Speech – Verb	-0.12	0.23	-0.57	0.35
Condition*Age Interaction	0.45	0.43	-0.39	1.31

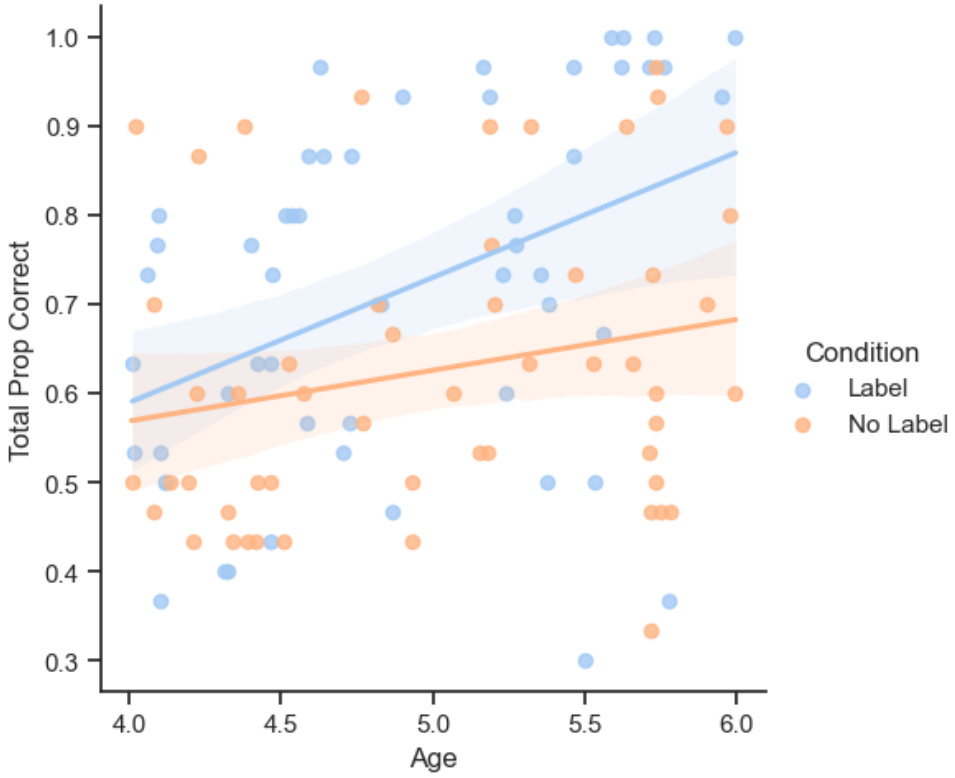


Figure 19. Proportion accuracy across all parts of speech tested in the pictorial analogy task as a function of age, separated by condition.

Overall, there were no significant differences in how four-year-old children performed across parts of speech, regardless of condition. In the no-label condition, there was no significant difference between how four-year-old children performed on trials involving adjectives (M = .596, SD = .213) and nouns (M = .608, SD = .150) ($t(24)$ = -.336, p = .740) and no

significant difference in performance on verb trials (M = .536, SD = .220) compared to noun trials (M = .608, SD = .150), $t(24) = 1.869$, p = .074) (see Figure 20). Similarly, there were no differences between how four-year-olds in the no-label condition performed on adjective and verb trials, $t(24) = 1.455$, p = .159.

In the label condition, there was no significant difference between how four-year-old children performed on trials involving adjectives (M = .666, SD = .229) and nouns (M = .645, SD = .190) ($t(28) = .606$, p = .550) and no significant difference in performance on verb trials (M = .672, SD = .191) compared to noun trials (M = .645, SD = .190), $t(28) = -.731$, p = .471) (see Figure 20). Similarly, there were no differences between how four-year-olds in the no-label condition performed on adjective and verb trials, $t(28) = -.203$, p = .841.

When examining what drives the label effect for five-year-olds, we found that five-year-olds in the label condition performed significantly better on adjective trials (M=.808, SD=.232) and noun trials (M=.754, SD=.238) than those in the no label condition (M=.707, SD=.181 and M=.628, SD=.217, respectively) ($t(51) = -1.787$, p = .04 and $t(51) = -2.025$, p = .024, respectively).

Collapsed across conditions, five-year-old children performed significantly more accurately on trials involving adjectives (M = .753, SD = .2099) than nouns (M = .685, SD = 233) ($t(52) = 2.590$, p = .012) and performed overall higher on verb trials (M = .736, SD = .226) compared to noun trials (M = .685, SD = 233), though there were no significant differences between the two ($t(52) = -1.895$, p = .064). Similarly, there were no differences between how five-year-olds performed on adjective and verb trials ($t(52) = .894$, p = .376).

Splitting it up by conditions, we see differences across parts of speech specifically for children in the no-label condition. In the no-label condition, five-year-old children performed

significantly more accurately on trials involving adjectives (M = .707, SD = .181) than nouns (M = .628, SD = .217) (t(28) = 2.484, p = .019) and performed overall higher on verb trials (M = .697, SD = .199) compared to noun trials (M = .628, SD = .217), though there were no significant differences between the two, t(28) = -1.808, p = .081 (see Figures 20, 21). Similarly, there were no differences between how five-year-olds performed on adjective and verb trials, t(28) = .341, p = .736.

In the label condition, there was no significant difference between how five-year-old children performed on trials involving adjectives (M = .808, SD = .232) and nouns (M = .754, SD = .238) (t(23) = 1.236, p = .229) and no significant difference in performance on verb trials (M = .783, SD = .251) compared to noun trials (M = .754, SD = .238), t(23) = -.771, p = .448) (see Figures

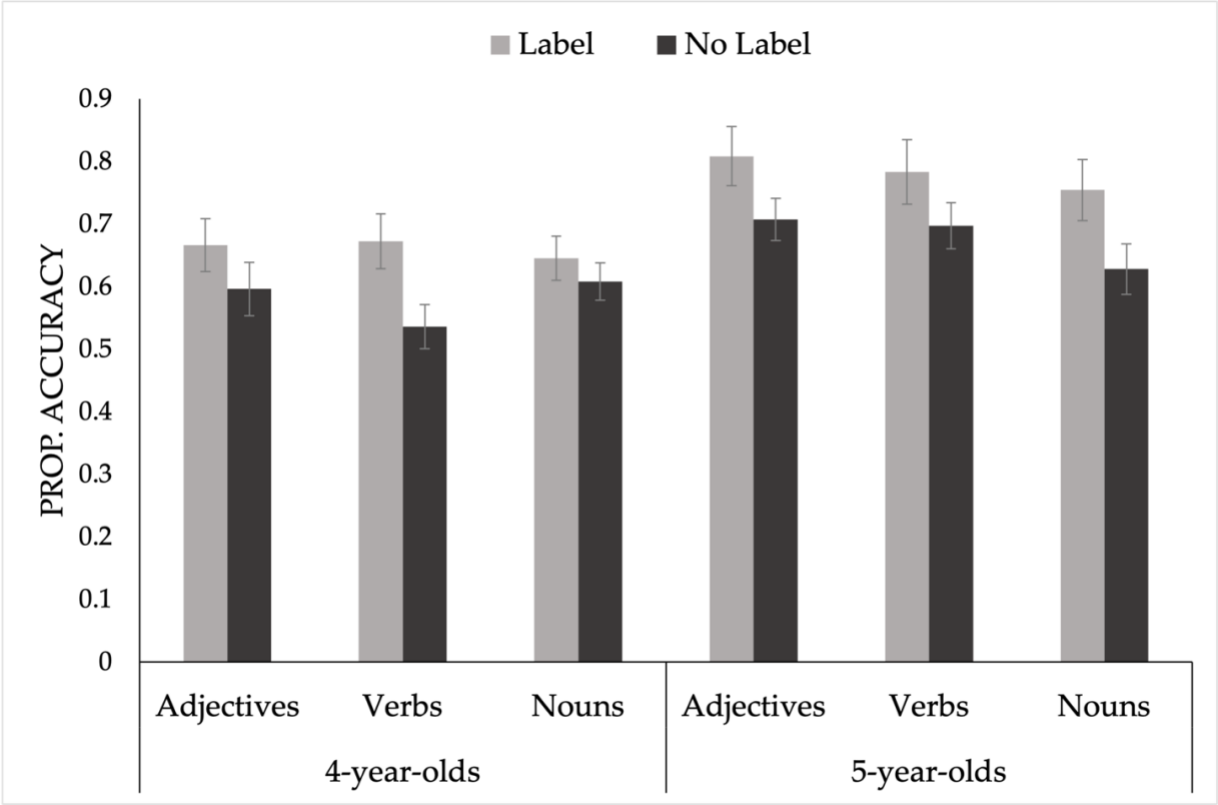


Figure 20. Average proportion accuracy for each condition across three lexical classes, separated by age group. Error bars reflect ± 1 standard error of the mean for human responses.

20, 21). Similarly, there were no differences between how five-year-olds in the label condition performed on adjective and verb trials, $t(23) = 1.187$, $p = .247$.

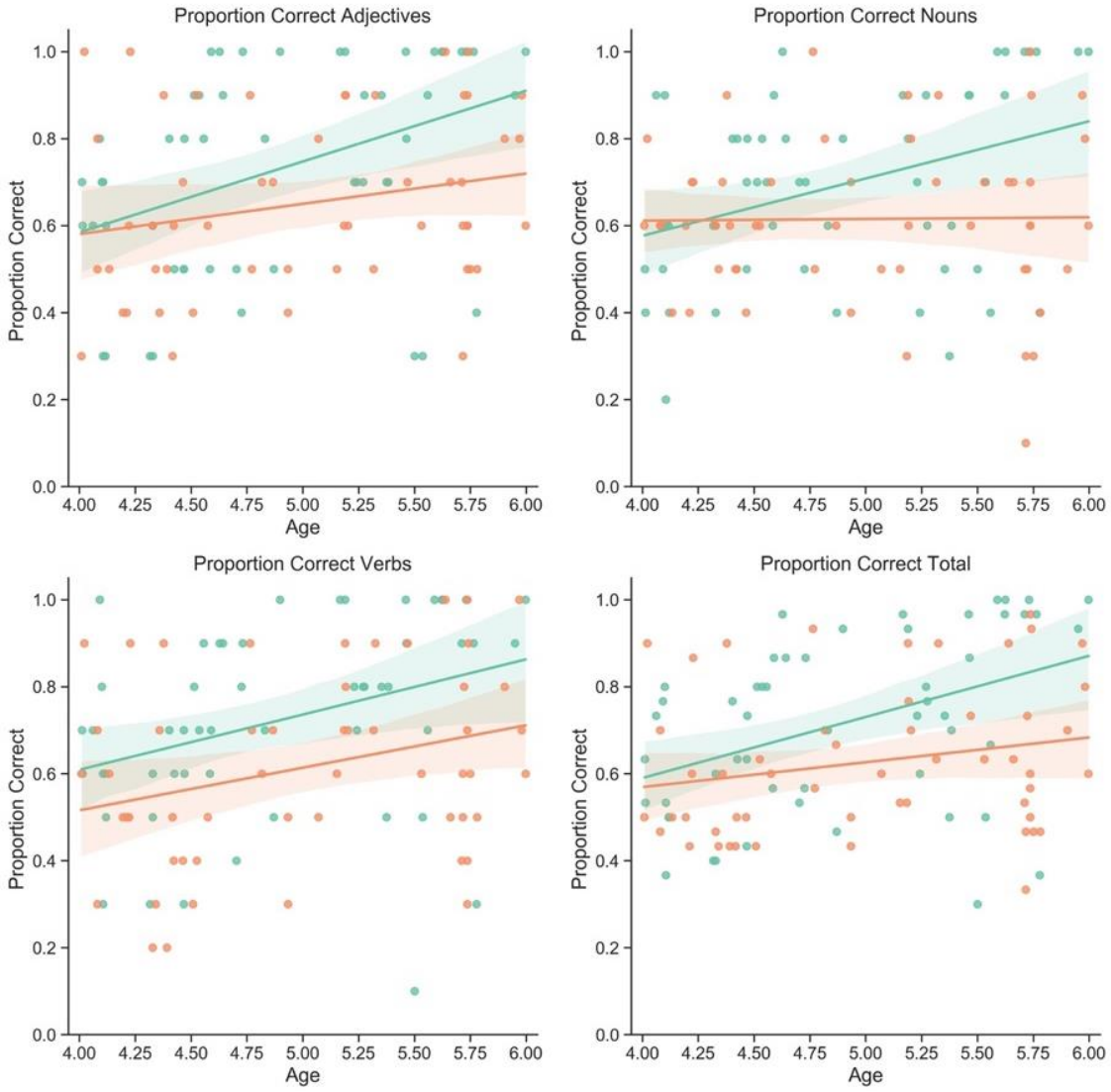


Figure 21. Response distribution for Label and No-Label conditions across each lexical class and overall, as a function of age. Green represents the label condition and orange represents the no-label condition.

IIB. Computational

We implemented two computational models of verbal analogy, Word2Vec (Mikolov et al., 2013) and BART (Lu et al., 2019), to compare model predictions with children's performance.

Both models operate on vector representations (*embeddings*) of individual word meanings. However, as shown in Figure 22, the two models operationalize in different representation spaces. Word2Vec is based on semantic space for individual words, such that words with similar meanings are clustered together in this semantic space. In contrast, BART forms relation space, in which each dimension indicates a specific relation. Hence, word pairs instantiating similar relations are located closely in the BART relation space. Based on representations of the two words in each pair, the models compute the dissimilarity of a source word pair with a target word pair and select the option with the smaller dissimilarity value as the predicted correct response.

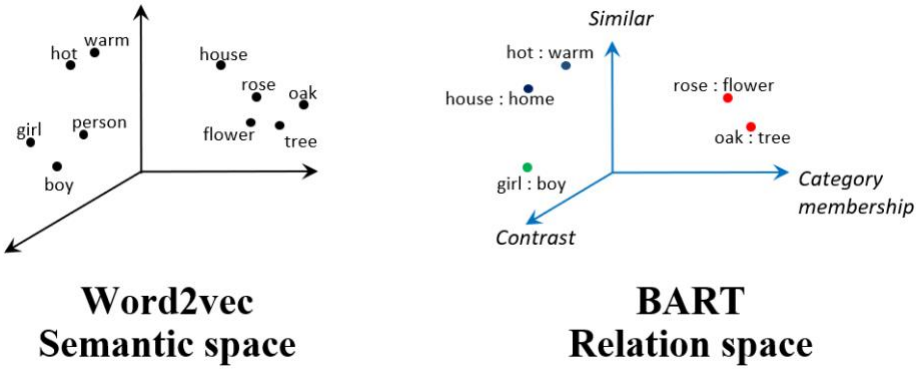


Figure 22. Illustration of Word2vec semantic space for individual words, and BART relation space for word pairs.

Word embeddings produced by Word2vec (Mikolov et al., 2013) were used to represent the meanings of each of the words included in the test trials of the pictorial analogy task (90 word pairs, with 180 total word embeddings). Word2vec-diff is a measure defined as the difference between the vectors of each word in a pair: i.e., $f_A - f_B$ for the word pair *A:B*. The dissimilarity between two pairs is then defined by the cosine distance between the difference vectors for the two pairs:

$$D_{W2V-diff} = cos\ (f_A - f_B,\ f_C - f_D)$$

The second model, BART, is trained on a set of specific relations, including 79 abstract relations from the SemEval-2012 Task-2 dataset (Jurgens et al., 2012) and an additional 56 relations (Popov et al., 2017). For each of those relations, BART was trained with less than 100 examples, including a small number (10 or 20) of positive examples instantiating this relation and some negative examples (~70) that instantiate other relations.

After learning explicit representations of each semantic relation, BART encodes the specific relation between any pair of words (A, B) using distributed representations expressed as a relation vector R_{AB}, in which each element indicates the probability that this pair of input words instantiates each of the learned relations. The relation vector is 270 dimensions (including the 135 relations in the training datasets and their corresponding converses). To solve an analogy problem, the model computes dissimilarity as the cosine distance between corresponding relation vectors based on the two word pairs and selects the answer with smaller dissimilarity:

$$D_{BART} = cos\ (R_{AB},\ R_{CD})$$

Additionally, this simulation was run on the NLP model Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). BERT is trained on massive text corpora and intended to predict words based on the highest probability that they would occur in a particular sentence (e.g., "A [x] is a type of bird."). In lieu of being trained on explicit relation representations, BERT relies solely on the frequency of word usage in their training corpora.

We conducted two simulations on BERT with the input "is related to" or "is opposite to," the latter of which is analogous to the condition of the behavioral task in which we provided children with a relational label. As such, we were able to compare its performance to the other two models when given the "is related to" input, as well as the to the behavioral task results from Experiment 2A in both conditions.

It is important to note that all three models received only the verbal input (e.g., happy:sad), unlike the children in the behavioral task, who also received accompanying pictures. Because some of the present-tense verbs in our task could also be interpreted as nouns, we ran simulations on all models on all verb tenses to compare whether it was the verb present tense that was perhaps hindering model performance.

Results

For both models, the dissimilarity between the word pairs was computed using the cosine distance between the vectors representing each pair. If the cosine distance between the source pair and the target pair was less than that between the source pair and the distractor pair, we considered that the models had correctly answered the analogy problem. Note that models W2vec-diff and BART are not sensitive to the presence of relation labels. Accordingly, we compared model predictions and children's performance in the No-Label conditions. However, BERT is able to receive input such as "is opposite to" and thus is able to be compared to children's performance in the label condition, as well.

Our simulations found that BART performed most accurately on adjective antonym pairs (.80 correct), followed by noun and verb pairs (.60 and .40, respectively (see Figure 23).

Word2vec also performed most accurately on adjective pairs (.70 correct), followed by verbs and nouns (.60 and .50, respectively). BART performed comparably to, or better than, children on adjectives and nouns but failed to match their performance on verb trials. Word2vec performed better than or comparably to children on adjectives and verbs but lower on nouns.

BERT was run on two separate simulations. In one, BERT was given the input "is related to" and it performed most accurately on noun pairs (.70 correct), followed by adjective and verb pairs (.60 and .50, respectively) (see Table 3). In another simulation, BERT was given the input "is opposite to," which serves as a better comparison to the condition in which children were given the label "these are opposites." In this simulation, BERT performed most accurately on verb pairs (.90 correct), followed by adjective and noun pairs (.70 and .60, respectively). These results are comparable to those of five-year-old children in the label condition, suggesting that such input plays a similarly beneficial role for both children and models such as BERT.

Overall, both models showed variability across the different parts of speech. Both models yielded levels of accuracy approximating (or higher than) that of five-year-olds in the No-Label condition for antonyms based on adjectives and nouns. However, for verb antonyms, the models (particularly BART) fell well short of the level achieved by five-year-olds.

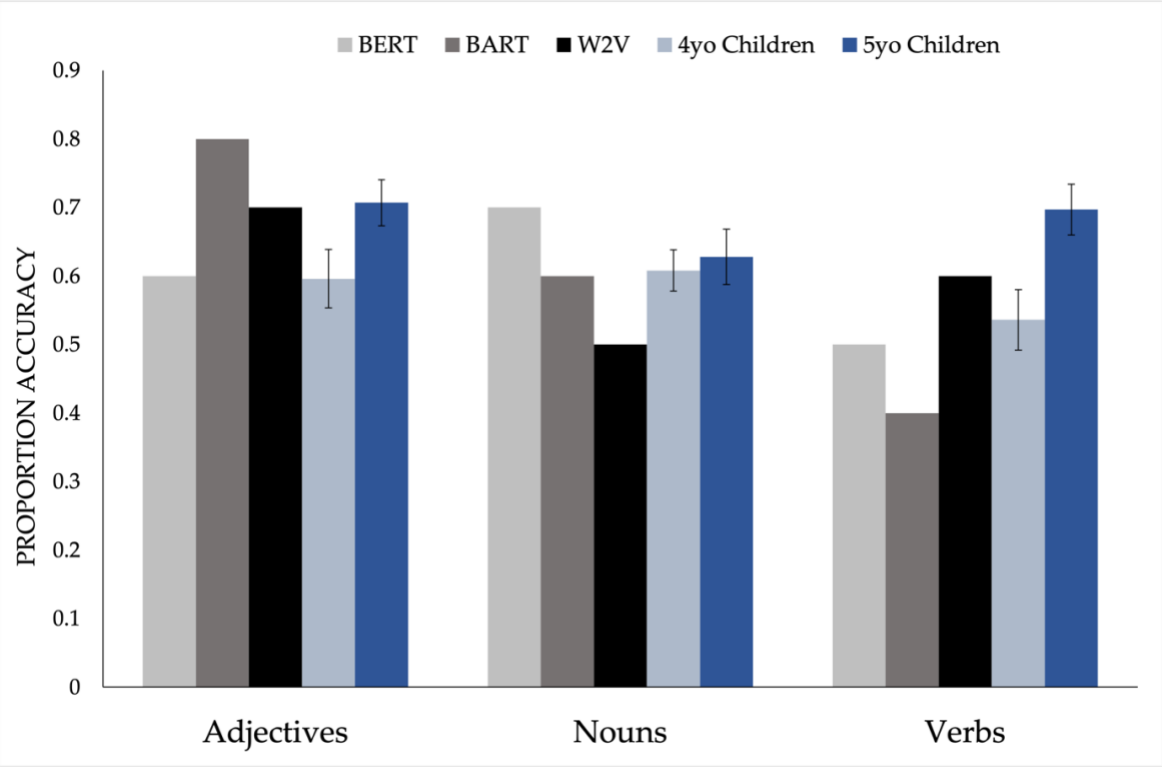


Figure 23. Model performance on the analogy task, and children's performance on the analogy task in the no-label condition.

Table 3. Model Performance Across Parts of Speech

	Models	Adjectives	Nouns	Verbs
Relation model	BART-270dim	0.80	0.60	0.40
NLP models	W2V	0.50	0.20	0.50
	BERT with input "is related to"	0.60	0.70	0.50
	BERT with input "is opposite to"	0.70	0.60	0.90

One of the potential explanations for why all models performed more poorly on verbs compared to adjectives and nouns (apart from BERT with the input of "is opposite to") could be that some of the verbs in the dataset could be mistaken for nouns while in the present tense (e.g., "love" or "frown"). To test this, we conducted simulations on each of the three models

using five tenses for each verb (present, past simple, present continuous, present perfect, and present-third). We found some variability in how the models performed across the different verb tenses. However, performance only increased for BART when verbs were provided in the present continuous tense (.70 correct) compared to the default present tense (.40 correct). Therefore, the variation in verb tenses does not appear to help model performance (see Table 4). Overall, taking the average performance across tenses, BERT performed similarly to five-year-old children when given the relational input "is opposite to." However, when given the input "is related to" (see Figure 24), BERT performed similarly to four-year-old children.

Table 4. Model performance across verb tenses.

Verb Tenses	**BART**	**BERT**	**W2V**
Present (e.g., throw)	0.40	0.89	0.50
Past Simple (e.g., threw)	0.50	0.67	0.30
Present Continuous (e.g., throwing)	0.70	0.89	0.20
Present Perfect (e.g., thrown)	0.50	0.67	0.30
Present-third (e.g., throws)	0.40	0.44	0.30
Average	**0.50**	**0.71**	**0.32**

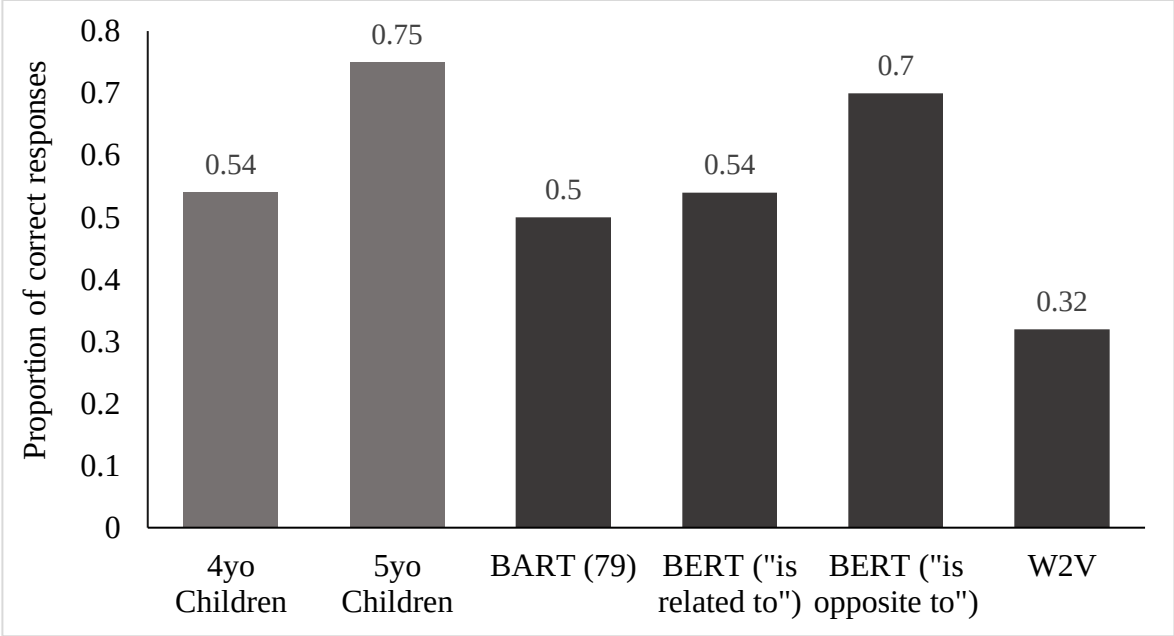


Figure 24. Proportion accuracy on the trials involving verb pairs for children and models.

Chapter II Discussion

Chapter II applied both developmental and computational methods to examine the solution of analogy problems based on antonyms. Using a verbal analogy task (with picture illustrations), we demonstrated that by age five years—before the antonym relation is formally taught in school—children are able to reliably solve analogies based on antonyms, especially when the antonym relation is given a verbal label ("opposites"). Namely, we find that children perform better on analogy problems involving antonyms when they are given a language cue in the form of a label than when they are faced with having to implicitly recognize the relation that the pairs hold.

In our experiment, we found no differences in performance across lexical classes for four-year-olds, who did not perform significantly above chance on any of the three. However, we found that five-year-olds, when not given a label for the relation, were more accurate on

analogy problems involving adjectives compared to nouns but performed comparably on problems involving verbs and those involving nouns.

Chapter I found that children could more reliably generate antonyms for nouns and adjectives than verbs. However, in Chapter II we find that problems involving nouns tend to be the hardest, specifically for children who are not provided relational language cues when solving analogy problems. These findings suggest that developmental trends in reasoning about the antonym relation do not coincide with children's lexical development or ability to generate pairs of words, given that children know more nouns than adjectives and verbs (Nelson, 1973; Sandhofer & Smith, 2007). A possible explanation is that nouns can be compared on a wider range of dimensions than either adjectives or verbs, making it more challenging to determine the basis for an antonymy relation for nouns.

We also compared children's performance to that of two vector-based models of verbal analogy, Word2vec (Mikolov et al., 2013) and BART (Lu et al., 2019), using the same set of problems. These models are based on embeddings derived from training on corpora of adult language. However, neither model is sensitive to the provision of verbal labels for the antonym relations, and thus, we compared their performance only to that of children in the no-label condition. BART specifically performed either better or comparably to children for adjective and noun trials but fell short for verb trials. Word2vec performed similarly to children for adjectives and verbs but worse for nouns.

In contrast, we also examined BART's performance against that of children and found that its performance is slightly lower for adjectives and verbs and higher on nouns than that of five-year-old children. We found that when BART is given the input "*is opposite to,"* BART

performs comparably to or higher than five-year-olds for adjectives, nouns, and verbs. This suggests that the type of language cues provided to children can also serve as input for NLP models such that the relation is explicitly stated, making these verbal analogies easier to complete across all parts of speech.

Variability in performance across parts of speech for both models and children suggests differences in the complexity of semantic meaning for each lexical class, which could play a role in both children's and models' ability to solve verbal analogies. While prior research on antonyms focused specifically on adjectives, this chapter shows that varying the lexical class of each analogy problem has potential implications for examining varying levels of performance on an identical task.

Chapter III: Analogies and Semantic Distance

Introduction

The experiments in this chapter sought to expand those in Chapter II. Experiment IIIA focused on examining the role of labels in helping children solve analogy problems involving the antonym relation *across* different lexical classes. Namely, do children perform differently for word pairs from different parts of speech when they can compare them within a single analogy problem? Moreover, do labels help ameliorate any differences? Unlike the experiments in Chapter II, which examined performance on analogy tasks involving a single part of speech, these experiments sought to compare performance on analogy problems involving both one part of speech individually (replicating experiment IIA) as well as multiple

within a single trial. Additionally, Experiment IIIB examined performance on the same analogy task on two models: BART and GPT3.5.

For Experiment IIIA, we predicted that, overall, children in the relational label condition would do better than children in the no label condition. For children in the label condition: within the trials involving the same part of speech, we predicted that children would do better on the adjective trials than the noun trials, in line with the findings from Chapter II showing that 5-year-old children perform better on analogy problems involving adjectives than those involving nouns when not given a label. Within the trials involving mixed parts of speech, we predicted that children would perform similarly on adjective and noun trials, though overall performance for both adjective and noun trials would be lower than that for the trials involving the same part of speech analogies. Though it is likely more difficult for children to draw analogies between pairs that are semantically different (adjectives vs. nouns), being able to compare the target and distractor noun pairs with each other and then, in turn, with the source adjective pair might help children detect how the pairs share the same abstract relation.

For the children in the no-label condition: within the trials involving the same part of speech, we predicted that there would be no differences between adjective trials and noun trials for either same vs. mixed analogy types.

IIIA. Behavioral

Methods

Participants

This experiment involved 113 participants, including 52 5-year-old (*M*=5.549, *SD*= 0.293) and 61 6-year-old (*M*=6.425, *SD*= 0.310) children recruited through UCLA's Language and Cognitive Developmental Lab. Only children whose parents provided consent participated, as required by the UCLA IRB. Estimated sample sizes were calculated using G*Power 3.1.9.4 (Faul, Erdfelder, Lang, & Buchner, 2007). Using Cohen's f of .2, a power analysis revealed that a sample size of 84 participants was required to achieve 95% power to detect (42 children per condition).

Design

This experiment involved a 2 (relational language) x 2 (analogy type) design. Relational language (relational label vs. no label) was manipulated between subjects to eliminate potential carry-over effects. Analogy type (same part of speech vs. mixed part of speech) was manipulated within subjects.

Parent Measures

As in the previous studies, parents completed a language survey. The language survey included all the words used in the study to determine 1) whether children have prior knowledge of the words used in the study and 2) whether word knowledge is related to their performance on the antonym task. In addition, parents were asked to complete a demographic questionnaire.

Stimuli

The experiment consisted of a pictorial analogy task that followed the format of a Relational Match-to-Sample (RMTS) task involving the same materials as the experiment in Chapter II. This format allowed children to compare the source pair simultaneously, exemplifying a contrast relation to the target pair that exemplifies a contrast relation but in the same or different lexical classes. For every trial, children were presented with a source pair, a target pair, and a distractor pair of semantically unrelated words.

In the *mixed* part of speech condition, if the source pair involved an adjective pair, the target pair might involve nouns (see Figure 25). The source pairs always had different parts of speech than the target and distractor pairs, whose parts of speech were always consistent. An example of the contrastive relations used in the task could be happy : sad (source adjective pair) :: friends : enemies (target noun pair), or friends : winners (semantically unrelated noun distractor pair).

In the *same* part of speech condition, the trials followed the same format except that the source pair, target pair, and distractor pair were always the same part of speech (see Figure 26). An example of the contrastive relations used in the task could be big : small (source adjective pair) :: sad : happy (target adjective pair), or sad : bored (semantically unrelated adjective distractor pair).

The pairs corresponding to antonyms and the distractor pairs were pictures of people, familiar animals, or objects.

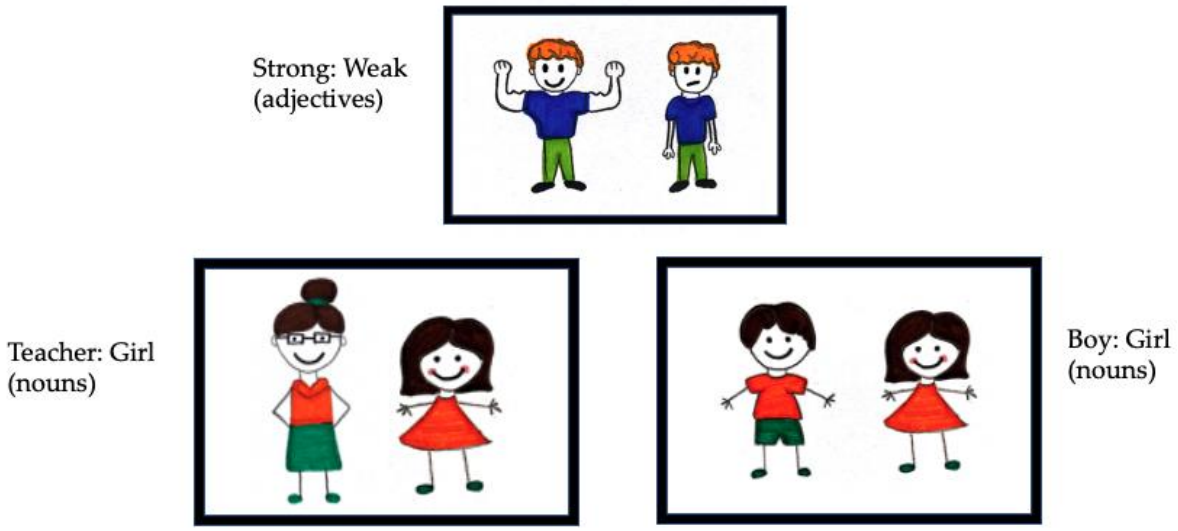


Figure 25. An example trial of the mixed part of speech condition. Source pair is at the top, distractor on the left and target pair on the right. The order of the target/distractor is randomized between trials.

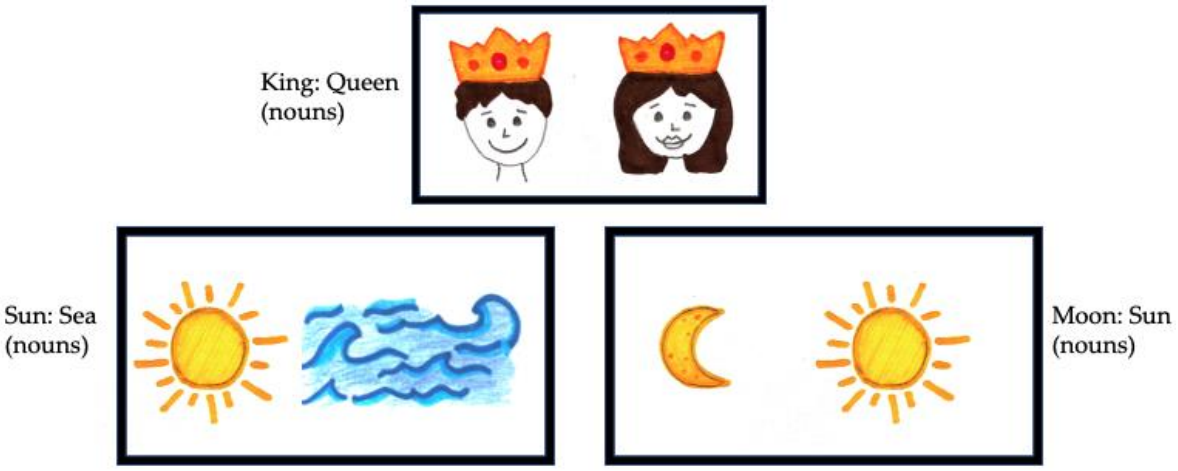


Figure 26. An example trial of the same part of speech condition. Source pair is at the top, distractor on the left, target on the right. The order of the target/distractor is randomized between trials.

Procedure

Children received two training trials and twenty test trials. Of the twenty trials, ten contained same part of speech pairs (five for adjectives and five for nouns), and the other ten contained mixed part of speech pairs (both nouns and adjectives; five in which the source pair contained adjectives and five in which the source pair contained nouns). Children were first

assigned to one of the two conditions: the Label and No-Label conditions. On each trial in the Label condition, children were told that the animals/objects/humans depicted in the source pair are "opposites" of each other (e.g., "This is dirty, this is clean. Dirty and clean are opposites") in both practice and test trials. The words used to describe objects were either adjectives or nouns. In the No-Label condition, children were not given a label for the abstract relation in any trials; instead, they were only given verbal descriptions of the individual objects (e.g., "This is dirty, and this is clean").

During practice trials, children were shown a source picture of two animals, objects, or humans depicting a pair of antonyms. The experimenter described the pictures, emphasizing the words that depicted the contrastive relation (e.g., "This is big, this is small. Big and small"). Afterward, the experimenter simultaneously provided two more images that respectively depict either a target pair of antonyms (e.g., "friends and enemies") or a distractor pair of semantically unrelated words (e.g., "friends and winners"). The participants were asked, "Which one is like this one (pointing to the source picture)?" Children were given feedback: either told that they were correct or told the correct answer if the child provided an incorrect one. The format of test trials was identical to that of training trials, except that the children were not given any feedback regarding their answers on each trial.

Two versions of the materials were created in which we randomized the order of the trials, the combinations of source and target/distractor pairs, and whether the target/distractor were a part of a mixed analogy (noun and adjective source/target combination) or a same analogy (only noun pairs or only adjective pairs). Because some pairs of words repeated, the versions were controlled such that a pair that appears as a source cannot appear as a target later

in the session. However, pairs that appeared as a target may appear as a source later in the version. Additionally, the display position of the target/distractor pair was randomized between trials such that the correct pair appeared on the left side of the screen in half of the trials and on the right side of the screen in the other half.

Results

Due to an error in counterbalancing, one of the adjective targets appeared in both the training and test trials. We removed the trial from our analyses.

Similarly to the experiment in Chapter IIA, we implemented a Bayesian logistic regression model using the R package brms (Burkner, 2018). We tested hypotheses by fitting a logistic regression model predicting responses on the analogy task based on the interaction between condition (Reference = Label) and analogy type (Reference = "same" trials). This model included group-level effects of subject and item and allowed for heterogeneity in the intercepts of the effects of condition and analogy type. The model also included a grouping of the item types into two parts of speech to analyze differences between analogies based on adjectives vs. nouns. For the prior distributions in our model, we used a uniform (i.e., uninformative) distribution for the main effects and interaction coefficient, and used a t(3,0,2.5) for the random intercepts and their standard deviation. Specified in brms syntax, the model is:

Response ~ Condition*Analogy Type + PartofSpeech +Age + (1| Subject) + (1| Item)

The model revealed that being given the relational label of "opposite" credibly predicted higher accuracy on the analogy task (b= 0.64, 95% CI [0.11, 1.20]) (see Table 5 and Figure 27).

However, the interaction between condition (label vs. no label) and the analogy type (same vs. mixed) did not credibly predict the responses on the analogy task (*b*= 0.06, 95% CI [-0.39, 0.51]).

Table 5. Estimates of Posteriors for Bayesian Logistic Regression Model

Population-Level Effects	Estimate	Est. Error	Lower 95% CI	Upper 95% CI
Intercept	-0.47	1.37	-3.15	2.25
Condition – Label	0.64	0.28	0.11	1.20
Analogy Type – Same	-0.26	0.16	-0.56	0.04
Part of Speech – Noun	-0.64	0.35	-1.36	0.02
Age	0.34	0.22	-0.10	0.78
Condition*Analogy Type Interaction	0.06	0.23	-0.39	0.51

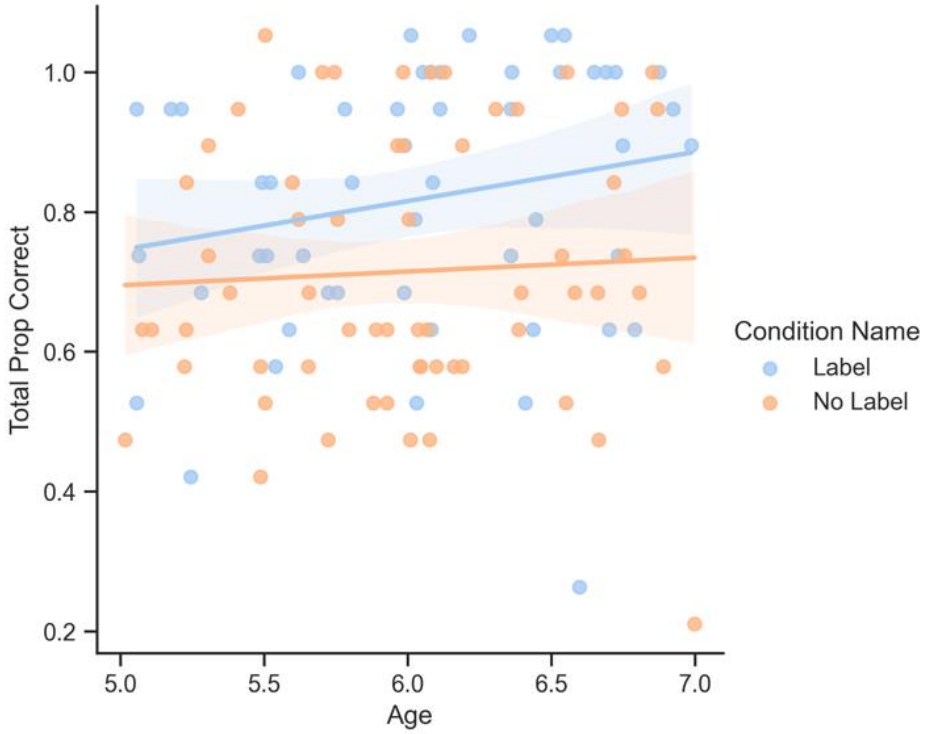


Figure 27. Proportion accuracy across both parts of speech tested in the pictorial analogy task as a function of age, separated by condition.

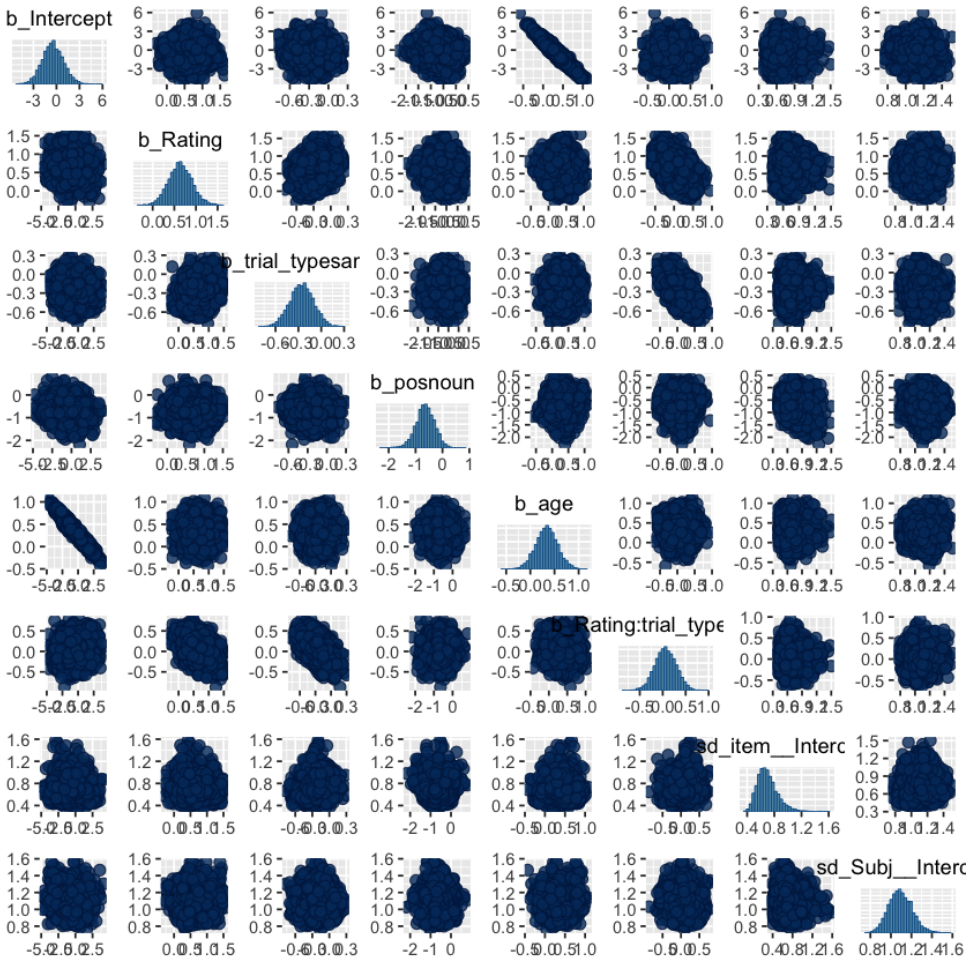


Figure 28. These plots represent correlations between the variables in our model with response as our dependent variable. "Rating" refers to condition, "trial_type" refers to "analogy type" and "pos" refers to "part of speech."

Looking at the correlations between our variables, we see that the intercept is highly correlated with age, suggesting that age increase has a strong linear effect on starting accuracy, despite the main effect of age not showing a credible effect on the responses (*b*= 0.34, 95% CI [-0.10, 0.78]) (see Figure 28).

We conducted an additional ANOVA to examine whether there are age differences between five- and six-year-olds for overall performance. We found that there were no differences in overall performance on the analogy task between five-year-olds (*M*= .743, *SD*=

.173) and six-year-olds (*M*= .778, *SD*= .210), *F*(1, 111) = .927, *p*=.338. There were also no differences in how they performed overall on trials involving the same part of speech, *F*(1,111)= .050, *p* = .823, and no differences in how they performed on trials involving mixed parts of speech, *F*(1,111) = 2.381, *p*= .126.

Furthermore, when examining performance strictly in the label condition, we found no significant differences in accuracy on same (*M*= .84, *SD*= .21) vs. mixed (*M*= .85, *SD*= .21) trials when the target pairs are adjectives, *t*(51) = -.337, *p*= .738 (see Figure 29). Similarly, there was also no significant difference between same (*M* = .719, *SD*= .25) and mixed (*M*= .785, *SD*= .23) trials when the target pairs were nouns, *t*(51)= -1.723, *p*= 0.91 (Figure 29).

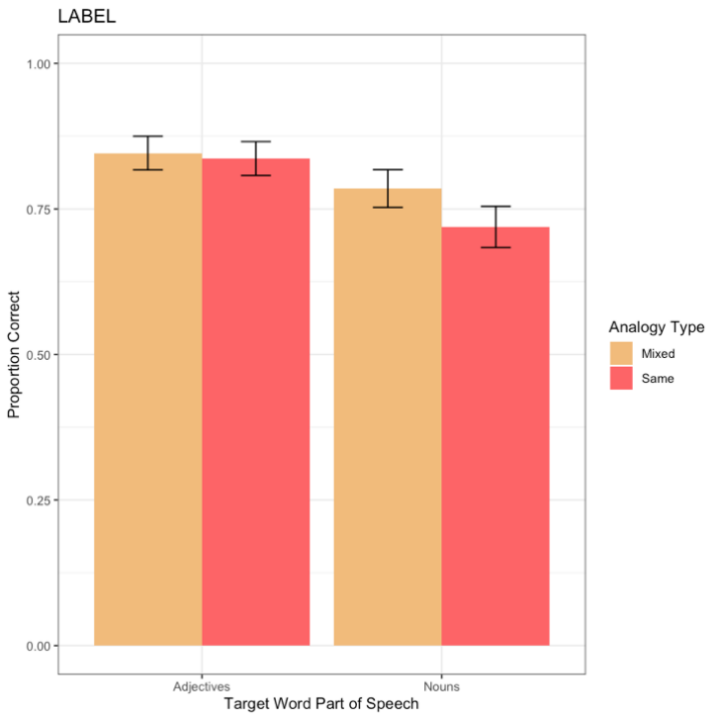


Figure 29. Proportion accuracy on same vs. mixed trials in the label condition. The x axis shows the part of speech that corresponds to the target words. The data is collapsed across ages. Error bars reflect ±1 standard error.

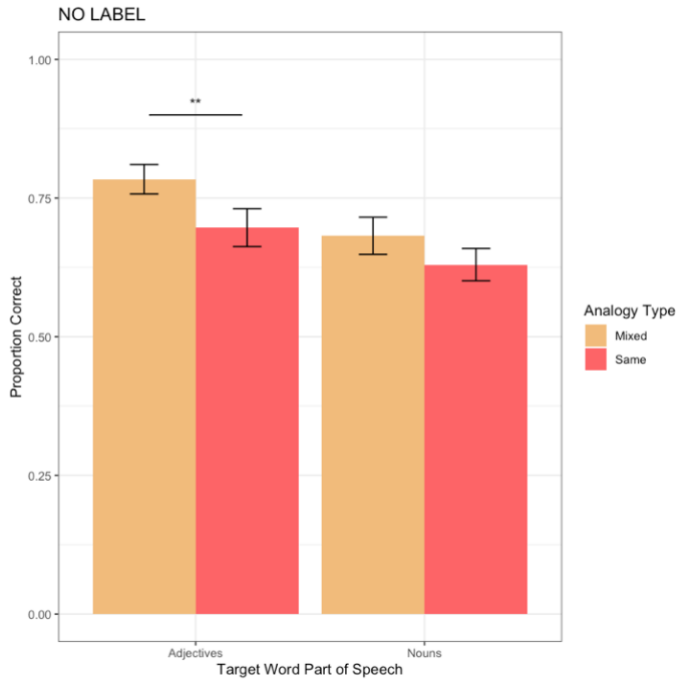


Figure 30. Proportion accuracy on same vs. mixed trials in the no-label condition. The x axis shows the part of speech that corresponds to the target words. The data is collapsed across ages. Error bars reflect ±1 standard error. ** represents a p-value lower than .01.

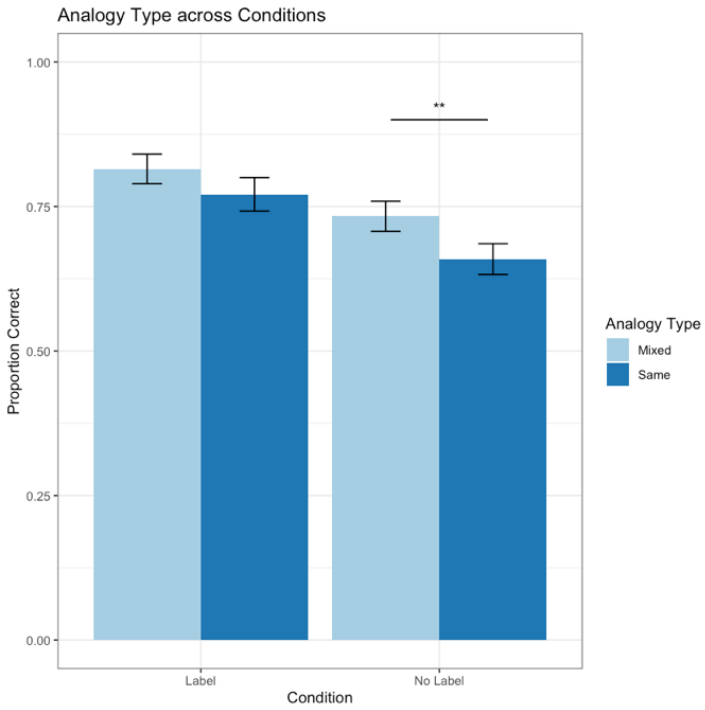


Figure 31. Proportion accuracy on same vs. mixed trials in both conditions. The data is collapsed across ages and part of speech. Error bars reflect ±1 standard error. ** represents a p-value lower than .01.

When examining performance strictly in the no-label condition, we found that accuracy on mixed trials in which the target pairs are adjectives (*M*= .784, *SD*= .208) is significantly higher than same trials (*M*= .697, *SD*= .267) in which the target pair is adjectives, *t*(60) = -3.167, *p*= .002 (see Figure 30). However, there was no significant difference between same (*M* = .630, *SD*= .228) and mixed (*M*= .682, *SD*= .262) trials when the target pairs were nouns, *t*(60)= -1.372, *p*= 0.175 (Figure 30).

When collapsing across the parts of speech, we found that performance on mixed trials (*M*= .733, *SD*= .204) is higher than that on same trials (*M*= .659, *SD*= .209) in the no-label condition, *t*(60)= -3.130, *p*=.003 (see Figure 31). However, in the label condition, we found no significant difference between mixed (*M*= .815, *SD*= .184) and same (*M*= .771, *SD*= .209) trials, *t*(51) = -1.759, *p*= .085 (see Figure 31).

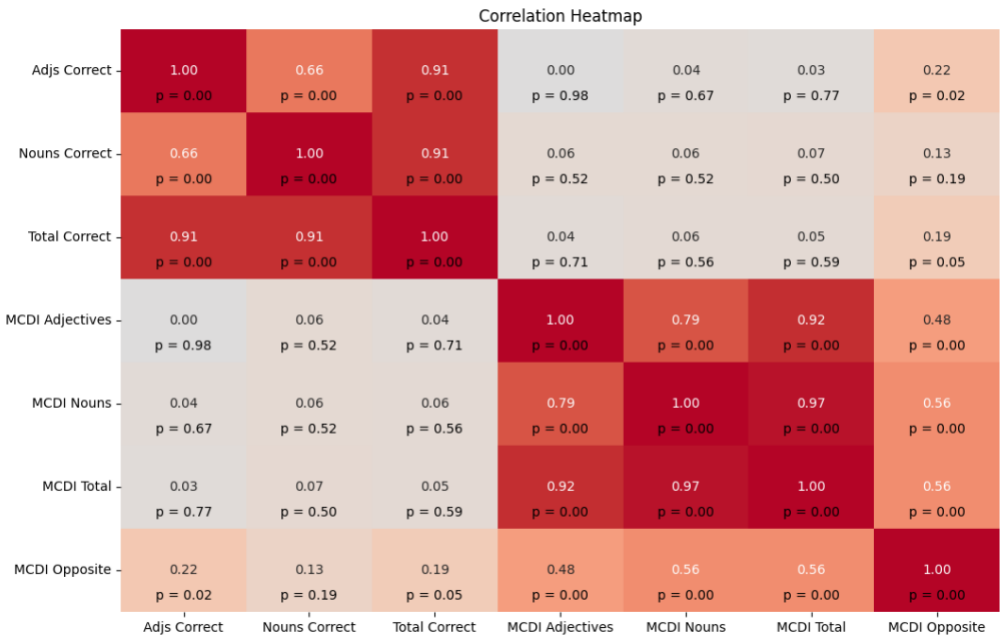


Figure 32. Correlations between accuracy on the antonym analogy task and parent reports of children's word knowledge. The accuracy on the task is shown as a total, as well as broken down by part of speech. Parent reported language ability is reflected by the language survey label. "MCDI Opposite" reflects parents' reports of whether children know the word "opposite."

We also conducted correlations between parent reports of children's language knowledge involving the words specifically used in our experiment and performance on the antonym analogy task. We found that overall performance on the antonym generation task and the knowledge of the words we expected children to know in order to complete this task was not significantly correlated $r(111)= 0.05$, $p=0.59$. However, there was a significant correlation between overall performance on the task and whether the parents reported if the children knew the word "opposite" $r(111)= 0.19$, $p=0.05$ (see Figure 32).

IIIB. Computational

In the computational portion of the final chapter, we implemented BART and GPT3.5. We administered the same task to GPT3.5 using both the no-label condition and the label condition. In the label condition, similar to children, the source pair was identified as depicting "opposites." In the no label condition, GPT3.5 was simply given the word pair, then asked which of the subsequent pairs (target and distractor) are like the source pair. GPT3.5 was given the verbal (written) input as that given to children, but it did not receive any visual input.

Results

We conducted the same task on BART. In our results, we included all the trials since BART did not need to be given training trials and thus the counterbalancing error in the behavioral data did not apply. We found that while BART performed similarly on trials involving both same and mixed analogies, GPT3.5 showed a similar performance on both

analogy types when given a label for the relation, but considerably better on the mixed trials when not given a label for the relation (see Table 6).

Table 6. Model performance on the antonym analogy task, separated by analogy type.

	Models	Same	Mixed
Relation model	BART-270dim	0.70	0.70
LLM model	GPT3.5 (no label)	0.70	0.90
	GPT 3.5 (label)	0.95	1

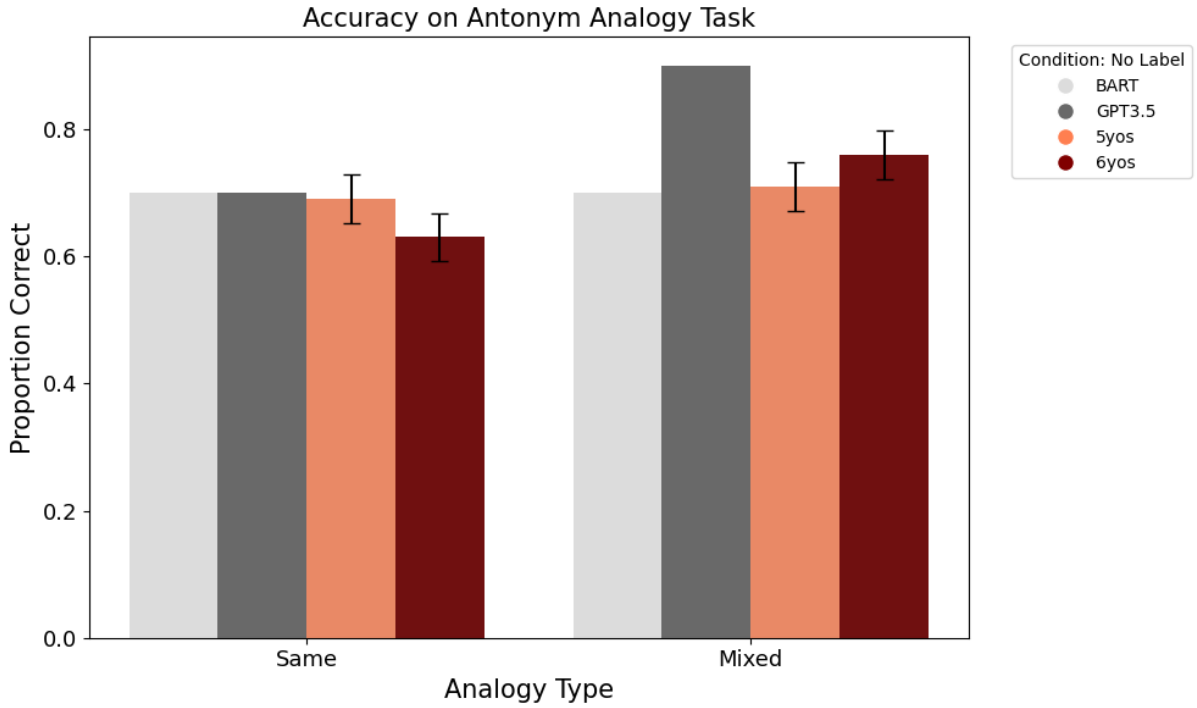


Figure 33. Model and human accuracy on the antonym analog task in the "no label" condition, separated by analogy type.

When comparing the performance of BART and GPT3.5 simulations to the performance of children in the no-label condition, we see that children perform comparably to the models for

both same- and mixed trials (see Figure 33). However, when given the label, GPT3.5 performs nearly perfectly on the task, outperforming children.

Chapter III Discussion

Chapter III served as an extension of Chapter II because we were interested in examining whether mixing the parts of speech within an analogy trial would lead to a difference in performance as opposed to only receiving trials in which there is only one lexical class for both the source the target/distractor. In this chapter, we also targeted a more advanced age range than the previous chapter: We tested five- and six-year-old children instead of four-year-olds due to at-chance performance on the verbal analogy task for four-year-olds in Chapter II. The experiments in Chapter III replicated those in Chapter II, namely by examining whether children can solve analogies in which each trial involves only one part of speech and manipulating whether language cues in the form of labels helped performance on these verbal analogies. However, we also presented children with trials in which the verbal analogy involved mixed parts of speech, namely, nouns and adjectives. Thus, the central goal of this chapter was to examine differences in children's and models' performance on verbal analogies that involved both the same part of speech within a trial as well as mixed parts of speech all in one trial.

In this chapter, we replicated Chapter II's findings, showing that giving children the label "opposite" when solving verbal analogies facilitates their performance on the analogy problems. This suggests that even as children's language abilities develop with age, providing a language cue to represent an abstract relation enhances their ability to detect the relation and

solve verbal analogies involving the relation. More surprisingly, we did not find significant age differences in this experiment, suggesting that although six-year-olds' performance on the analogy task is higher than that of five-year-olds, the differences are not significant, unlike those between four- and five-year-olds.

When looking at differences between trials involving the same part of speech and mixed parts of speech, we did not find differences between the same and mixed trials in the label condition. However, we found that performance on the mixed trials was significantly higher in the no-label condition compared to the same part of speech trials. These findings could be due to the fact that when not given a label, children have to rely on the comparison between source and target/distractor, and it could be that having the pairs contain different parts of speech highlights the abstract relation that the individual pairs share. These findings correspond to previous research showing that even by second grade, children are still susceptible to incorrectly categorizing two words as antonyms strictly because those words share a close association (e.g., salt and sea) (Landis et al., 1987). Similarly, GPT3.5 showed the same patterns of responses, suggesting that when both children and models are required to draw analogies based on the word pairs alone, without a relational language cue, it is easier for them to draw the analogy between word pairs that are less semantically related than those that belong to the same lexical classes.

Furthermore, because BART's performance is the same between same and mixed trials, this suggests that the pairs chosen reliably share the antonym relation. Though we had predicted that words from different lexical classes would have greater semantic distance between them, both combinations of problems (same part of speech vs. mixed parts of speech)

reliably form a strong relational vector with equal distances between the source and target relation vectors.

General Discussion

The goal of Chapter I was to examine how children generate antonyms and the types of antonym responses they generate. This chapter also sought to compare children's performance with relational models such as Bart-Gen on generating the antonym pair of a given query word. We sought to examine patterns in the generated responses of both children and the models in order to examine whether the generated words were semantically related to the target words even if they were not explicitly the target word. Overall, we found no age effect between 4- and 5-year-olds on the antonym generative task. We found differences in each age group's ability to produce antonyms across different parts of speech. Namely, we saw that children of both ages had an easier time generating antonym pairs when the query and target words were adjectives and nouns as opposed to verbs.

Additionally, we found that children's success on the antonym generative task was significantly correlated with their knowledge of the individual words. These findings show that children as young as four already have a surprising ability to generate various instantiations of the antonym relation, even when not presented with contextual aids, such as pictures. While previous work also found success with four- and five-year-olds' ability to produce antonyms, they relied on pictures to facilitate this performance (Phillips & Pexman, 2015). Previous work has also focused exclusively on adjective pairs, and this chapter expands our understanding of how lexical class knowledge interacts with children's ability to produce antonyms.

Our findings in Chapter I also demonstrate that semantic clusters are formed for adjective and noun responses more so than for verb responses. This suggests that the ability to use adjectives, nouns, and verbs in order to generate antonyms is consistent with patterns of lexical development that typically developing children tend to follow. Namely, children tend to learn adjectives and nouns first, followed by verbs, which coincides with our data showing that verbs pose a particular challenge for children solving the antonym task. Thus, older children may be able to reliably generate antonym pairs across all parts of speech once they gain more advanced semantic familiarity with all the words that the task requires them to know.

Chapter II aimed to expand on the experiment from Chapter I and examine how children can use pairs of antonyms to solve analogy problems. This study examined analogy problems involving the same three parts of speech from Chapter I. We aimed to examine whether giving the children relational language cues in labels such as the word "opposite" would help them solve verbal analogies involving antonyms. Additionally, we created a pictorial task so that children could see depictions of the words in the verbal task, reducing the cognitive load of remembering all the words from the source, target, and distractor for these preliterate children. We found that, indeed, language cues did facilitate performance on the analogy task, particularly for older kids. Relational labels make abstract relations more concrete and thus highlight the exact relation children are expected to detect from the word pairs.

Additionally, we found that children performed similarly to or better than the models even when not given a language cue. In contrast, we did not find significant differences between different parts of speech, though children performed the lowest on trials involving nouns. The ability to match a target pair to the source pair in these types of verbal analogies

becomes commonplace for adults. Therefore, examining the developmental origins of when and how children begin to reason about verbal analogies informs the processes involved in making relational analogies. Furthermore, this study provides an additional investigation into the role of relational language in facilitating reasoning about these abstract relations before children are formally taught about them in school and before they begin to use them in their daily lives.

Chapter III aimed to examine how children use the antonym relation to reason by analogy, both when the verbal analogy is between pairs of words that belong to the same lexical class and when the words belong to different lexical classes. This chapter was intended to serve as both a replication of the previous chapter through the trials that involved the same part of speech and to examine potential differences in performance between trials involving the same part of speech vs mixed parts of speech (nouns and adjectives combined in one trial). In this chapter, we also examined how language cues affect performance for comparably older children than in the previous chapter (5- and 6-year-olds compared to 4- and 5-year-olds in Chapter II). We also compared children's performance on various models in order to compute whether the potential increased semantic distance between different parts of speech would hurt performance for both children and models when solving these analogy problems.

We replicated the findings from Chapter II, showing that relational language cues facilitate performance for children. Furthermore, we found an effect of analogy type, in which children who were not given a label performed better on trials involving mixed parts of speech than those involving the same part of speech for both source and target/distractor pairs. These findings suggest that children benefit from comparing pairs of words belonging to different lexical classes when solving analogy problems without relational language cues. Given BART's

comparable performance on same vs. mixed trails, we conclude that the pairs of words that were chosen reliably demonstrate the antonym relation and thus lower the importance of semantic distance between the individual words that compose the pairs, making the task have similar levels of difficulty regardless of the type of analogy problem they are given.

These experiments demonstrate that children can successfully generate and use antonyms to reason by analogy long before formally being taught about them in school. This body of work provides evidence for how children can first grasp these abstract relations and produce them in various instantiations. While there is evidence showing that an understanding of antonyms, specifically, develops before other semantic relations (Landis et al., 1987; Heidenheimer, 1978), previous work suggests that this understanding is fully formed in middle school. Here, we examine the first emergence of this ability, and the types of responses children produce when completing an antonym generative task. Moreover, we were able to examine how current models fare against the ability of young children and how different parts of speech might play a role in their ability to solve these tasks successfully.

Beyond the generative task, Chapters II and III used a novel task that incorporates the typical verbal analogies that adults would be given on semantic relation analogy tasks and made it potentially more accessible for young children by including pictures that represent each of the words in the verbal analogy and thus acted as an anchor for what each pair represents, reducing the cognitive load required of the task. While the analogy task remained difficult for four-year-olds, we found that five- and six-year-olds showed high performance, especially when given relational language cues in the form of a label. We found the same for models like BERT, in which performance improves when provided with input that facilitates its search for

the correct relation. The findings from these chapters suggest that children can successfully draw analogies using the antonym relation even long before they are formally taught to do so, and their performance either matches or is better than that of models intended to match adult performance.

In conclusion, this series of studies highlights i) the importance of examining potential variability across different parts of speech in semantic relations and verbal analogies; ii) how relational cues can facilitate solving verbal analogies; iii) how varying lexical class within an analogy problem might highlight the relation for both children and models. Overall, these studies provide a comprehensive examination of the development of antonymy understanding in young children and identifies key factors that can enhance the ability of both children and models to solve verbal analogies.

Appendices

A. Pairs of words used in Chapter I. Adjectives are marked in green, nouns in blue, and verbs in orange.

Training Trials

QW1	QW2
clean	dirty
day	night
win	lose

Test Trials

QW1	QW2
big	little
cold	hot
dry	wet
fast	slow
good	bad
happy	sad
weak	strong
girl	boy
love	hate
outside	inside
queen	king
sun	moon
teacher	student
winter	summer
frown	smile
give	take
open	close
pull	push
throw	catch
walk	run
whisper	shout

B. Pairs of words used in Chapter II.

Training Trials

Target		**Target**		**Distractor**	
on	off	clean	dirty	clean	sad
top	bottom	day	night	day	summer
stop	go	win	lose	win	sit

Test Trials

Source		**Target**		**Distractor**	
strong	weak	cold	hot	cold	dirty
wet	dry	short	tall	short	angry
fast	slow	weak	strong	weak	happy
good	bad	empty	full	empty	small
happy	sad	fast	slow	fast	happy
cold	hot	dry	wet	dry	tired
empty	full	big	little	big	dirty
on	off	good	bad	good	wet
tall	short	awake	asleep	awake	sad
clean	dirty	happy	sad	happy	surprised
outside	inside	winter	summer	winter	night
teacher	student	queen	king	queen	woman
floor	ceiling	food	drink	food	plate
summer	winter	sun	moon	sun	sea
king	queen	friend	enemy	friend	mother
friend	enemy	girl	boy	girl	teacher
top	bottom	person	crowd	person	girl
boy	girl	teacher	student	teacher	mother
outside	inside	love	hate	love	smile
night	day	outside	inside	outside	city
sit	stand	throw	catch	throw	stop
open	close	frown	smile	frown	hate
cry	laugh	break	fix	break	open
work	play	build	destroy	build	stop
stop	go	whisper	shout	whisper	talk
win	lose	open	close	open	break
enter	exit	walk	run	walk	stand
give	take	pull	push	pull	stand
build	destroy	play	work	play	stop
throw	catch	give	take	give	close

C. Pairs of words used in Chapter III. Note that source/target pair combinations were randomized between versions, with each target/distractor used in both "same" and "mixed" analogies.

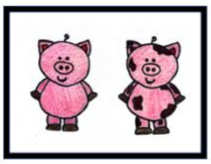
Training Trials

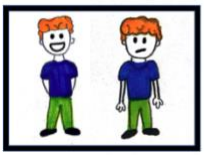
Source		Target		Distractor		Analogy Type	Target Part of Speech
big	little	clean	dirty	sad	dirty	same	adjective
big	little	night	day	day	summer	mixed	noun

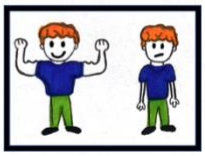
Test Trials

Source		Target		Distractor		Analogy Type	Target Part of Speech
tall	short	good	bad	good	wet	same	adjective
dry	wet	big	little	dirty	big	same	adjective
summer	winter	outside	inside	outside	city	same	noun
good	bad	fast	slow	happy	slow	same	adjective
fast	slow	happy	sad	surprised	sad	same	adjective
outside	inside	person	crowd	person	girl	same	noun
happy	sad	strong	weak	happy	weak	same	adjective
boy	girl	friend	enemy	friend	mother	same	noun
teacher	student	moon	sun	sun	sea	same	noun
friend	enemy	food	drink	food	plate	same	noun
cold	hot	king	queen	mother	queen	mixed	noun
outside	inside	cold	hot	strong	cold	mixed	adjective
king	queen	asleep	awake	sad	awake	mixed	adjective
clean	dirty	love	hate	hate	smile	mixed	noun
strong	weak	boy	girl	teacher	girl	mixed	noun
asleep	awake	teacher	student	mother	teacher	mixed	noun
top	bottom	tall	short	angry	short	mixed	adjective
floor	ceiling	dry	wet	tired	dry	mixed	adjective
night	day	empty	full	small	empty	mixed	adjective
empty	full	summer	winter	winter	night	mixed	noun

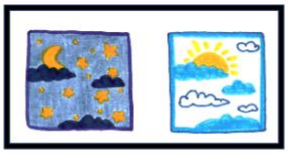
D. Pictures used in Chapters II and III. Note that the target/distractor remain together regardless of the trial (only the location might be flipped left/right), but the combination with the source pair is randomized across versions. These are examples of trials from one version of Chapter II. Chapter III did not include any verb pairs. The italicized pair represents the distractor.

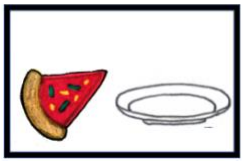


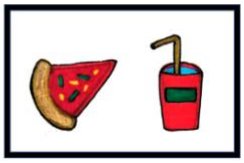




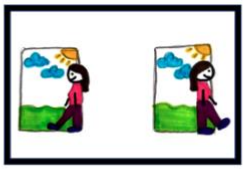
clean:dirty:: *happy:weak* or strong:weak

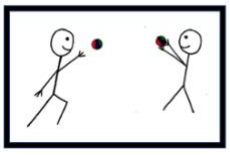


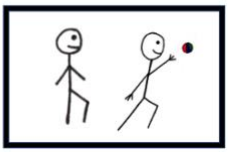




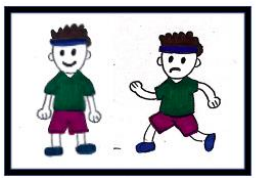
day:night:: *food:plate* or food:drink



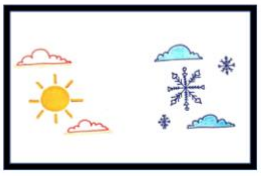




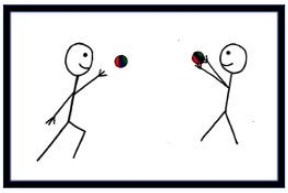
enter:exit::throw:catch or *stop:throw*

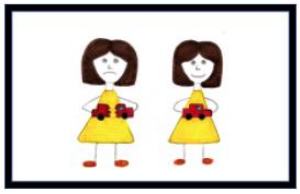


on:off:: *happy:slow* or fast:slow

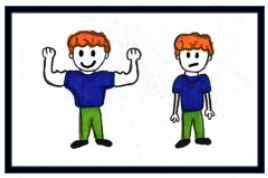


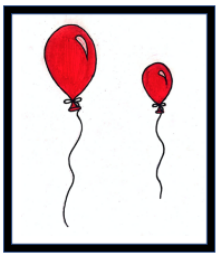
top:bottom:: *night:winter* or summer:winter

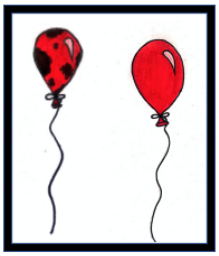




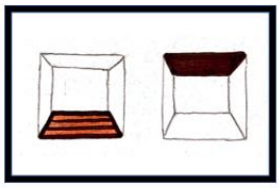
throw:catch:: *open:break* or break:fix

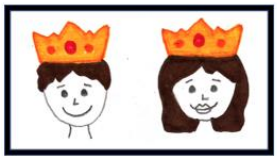




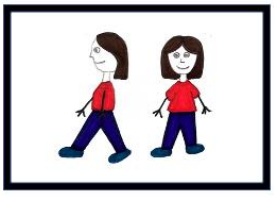


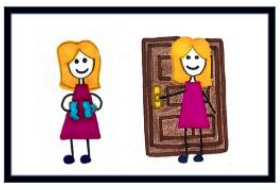
strong:weak::big:little or *dirty:big*



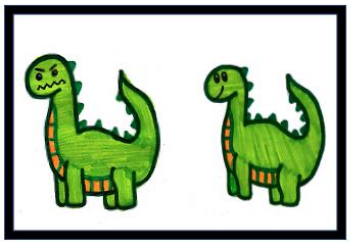


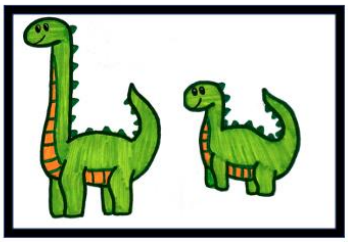
floor:ceiling::king:queen or *mother:queen*



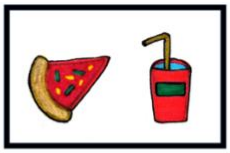


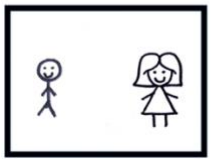
go:stop:: open:close or *break:close*

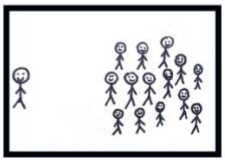




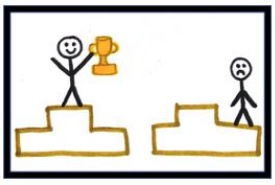
fast:slow::*angry:short* or tall:short

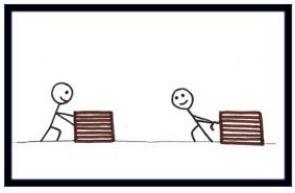


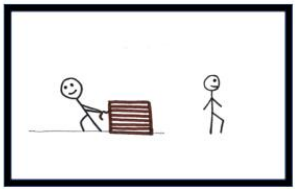




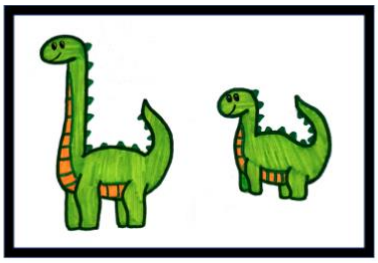
food:drink::*person:girl* or person:crowd

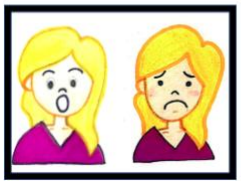




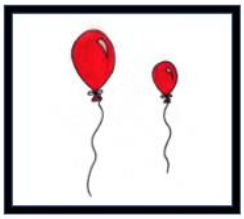


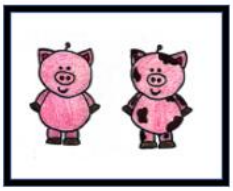
win:lose::push:pull or *pull:stop*

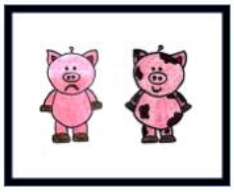




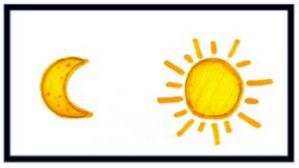
tall:short::*surprised:sad* or happy:sad



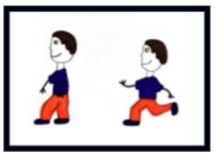


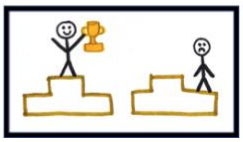


big:little::clean:dirty or *sad:dirty*

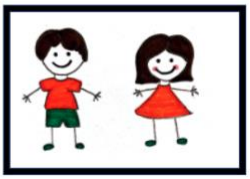


moon:sun::night:day or *day:summer*

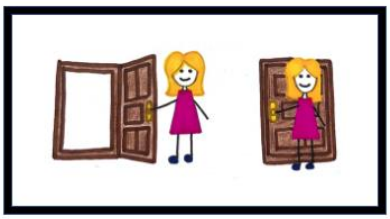


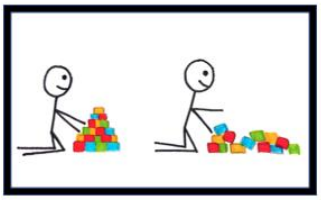


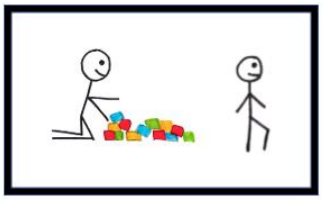
walk:run::*sit:lose* or win:lose



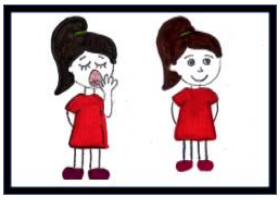
king:queen::*teacher:girl* or boy:girl





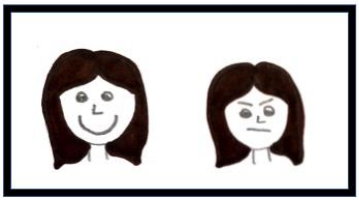


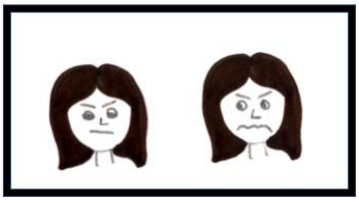
open:close:: build:destroy or *destroy:stop*



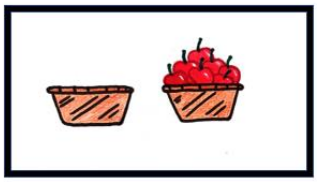
happy:sad::dry:wet or *tired:dry*

boy:girl::friends:enemies or *friends:mother*



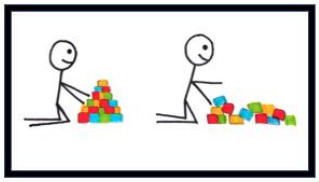


cry:laugh:: smile:frown or *frown:hate*

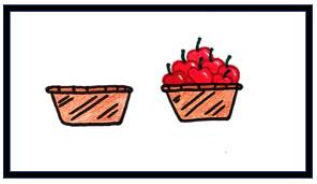


dry:wet::*small:empty* or empty:full

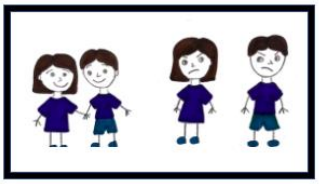
summer:winter:: love:hate or *hate:smile*

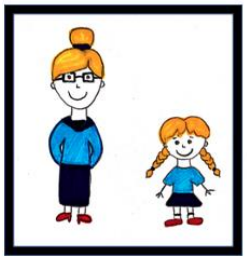


build:destroy::play:work or *play:stop*

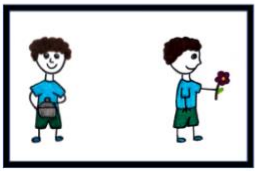


empty:full:: cold:hot or *strong:cold*

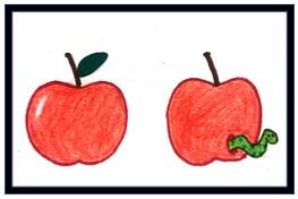


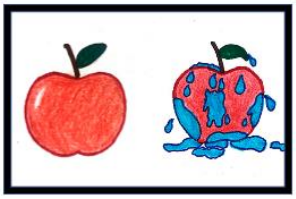


friends:enemies::teacher:student or *mother:teacher*

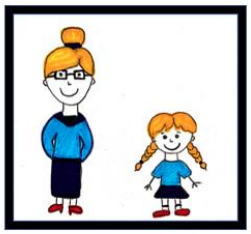


play:work::*close:give* or give:take

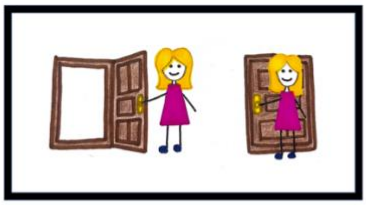


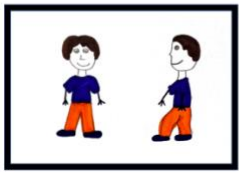


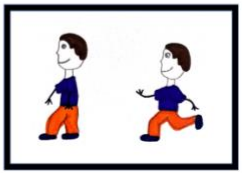
cold:hot:: good:bad or *good:wet*



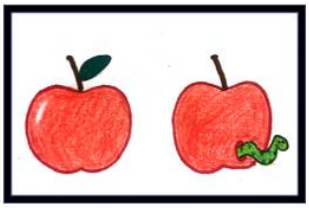
teacher:student::*outside:city* or outside:inside

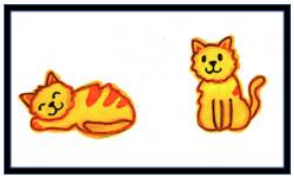


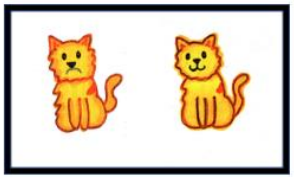




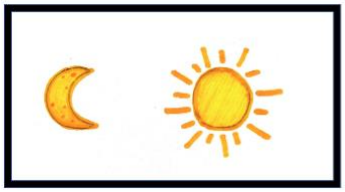
open:close::*stand:walk* or walk:run



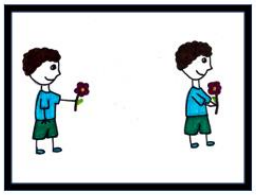


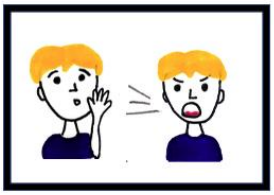


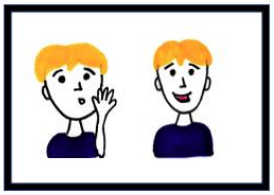
good:bad::asleep:awake or *sad:awake*



outside:inside::moon:sun or *sun:sea*







give:take::whisper:shout or *whisper:talk*

References

Anggoro, F., Gentner, D., & Klibanoff, R. (2005). How to go from nest to home: Children's Learning of abstract relational categories. In B.G. Bara & G. Salvendy (Eds.) *Proceedings of the 27th meeting of the Cognitive Science Society (*pp. 133-138). Baltimore: University Park Press.

Ankowski, A. A., Vlach, H. A., & Sandhofer, C. M. (2013). Comparison vs. Contrast: Task specifics affect category acquisition. *Infant and Child Development*, *22*, 1-23. doi: 10.1002/icd.1764

Asmuth, J. A., & Gentner, D. (2005). Context sensitivity of relational nouns. In B.G. Bara & G. Salvendy (Eds.) *Proceedings of the 27th Annual Meeting of the Cognitive Science Society (*pp. 133-138). Baltimore: University Park Press.

Baldwin, M. W. (1992). Relational schemas and the processing of social information. *Psychological Bulletin, 112*(3), 461. https://psycnet.apa.org/doi/10.1037/0033-2909.112.3.461.

Bowerman, M. (1976). Semantic factors in the acquisition of rules for word use and sentence construction. In D. Morehead, & A. Morehead (Eds.) *Directions in Normal and Deficient Language Development* (pp. 99-179). University Park Press.

Common Core State Standards Initiative (2017) English language arts standards (language, grade 4). Available at www.corestandards.org/ELA-Literacy/L/4/5/c/. Accessed August 11, 2017.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal*

Learning and Verbal Behavior, 8(2), 240-247. https://psycnet.apa.org/doi/10.1016/S0022-5371(69)80069-1.

Deneault, J., & Ricard, M. (2006). The Assessment of Children's Understanding of Inclusion Relations: Transitivity, Asymmetry, and Quantification. *Journal of Cognition and Development*, *7*(4), 551–570. https://doi.org/10.1207/s15327647jcd0704_6.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805.

Entwisle, D. R., Forsyth, D. F., & Muuss, R. (1964). The syntactic-paradigmatic shift in children's word associations. *Journal of Verbal Learning and Verbal Behavior, 3*(1), 19-29. https://psycnet.apa.org/doi/10.1016/S0022-5371(64)80055-4.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146.

Garnham, W. A., Brooks, J., Garnham, A., & Ostenfeld, A. M. (2000). From synonyms to homonyms: exploring the role of metarepresentation in language understanding. *Developmental Science, 3*(4), 428-441. https://doi.org/10.1111/1467-7687.00137.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology, 40*(2), 177. https://psycnet.apa.org/doi/10.1037/0012-1649.40.2.177.

Gentner, D. (1975). Evidence for the psychological reality of semantic components: The verbs of

possession. *Explorations in Cognition*, *35*, 211-246.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155-170. https://doi.org/10.1016/S0364-0213(83)80009-3.

Gentner, D. (2005). The development of relational category knowledge. *In Building object Categories in Developmental Time* (pp. 263-294). Psychology Press.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek, & R. Golinkoff, (Eds.) *Action meets word: How children learn verbs*, (pp 544–564). Oxford University Press. https://psycnet.apa.org/doi/10.1093/acprof:oso/9780195170009.003.0022

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*(5), 752-775. https://psycnet.apa.org/doi/10.1111/j.1551-6709.2010.01114.x.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition, 65*(2-3), 263-297. https://psycnet.apa.org/doi/10.1016/S0010-0277(98)00002-X

Gentner, D., & Namy, L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*, 487-513. https://doi.org/10.1016/S0885-2014(99)00016-7.

Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. *Child Development*, *82*(4), 1173-1188. https://psycnet.apa.org/doi/10.1111/j.1467-8624.2011.01599.x.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), Perspectives on language and thought: Interrelations in development (pp. 225–277). Cambridge University Press. https://doi.org/10.1017/CBO9780511983689.008

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology 15*(1): 1-38. http://hdl.handle.net/2027.42/25331

Goswami, U., & Brown, A. L. (1990). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition, 35*(1), 69-95. https://psycnet.apa.org/doi/10.1016/0010-0277(90)90037-K

Halford, G. S. (1993). *Children's understanding: The development of mental models.* Hillsdale, NJ: Erlbaum. https://doi.org/10.4324/9781315801803.

Hall, D. G., Waxman, S. R., & Hurwitz, W. M. (1993). How two-and four-year-old children interpret adjectives and count nouns. *Child Development, 64*(6), 1651-1664. https://psycnet.apa.org/doi/10.2307/1131461.

Heidenheimer, P. (1978). Logical relations in the semantic processing of children between six and ten: Emergence of antonym and synonym categorization. *Child Development*, *49*, 1243-1246. https://psycnet.apa.org/doi/10.2307/1128770.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*(3), 295-355. https://psycnet.apa.org/doi/10.1207/s15516709cog1303_1.

Ichien, N., Kan, A., Holyoak, K. J., & Lu, H. (2022). Generative inferences in relational and analogical reasoning: A comparison of computational models. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition, 62*(2), 169-200. https://psycnet.apa.org/doi/10.1016/S0010-0277(96)00784-6.

Jones, S., & Murphy, M. L. (2005). Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics*, *10*(3), 401-422. https://doi.org/10.1075/ijcl.10.3.06jon.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*(6), 2797-2822.

Landis, T. Y., Herrmann, D. J., & Chaffin, R. (1987). Developmental differences in the comprehension of semantic relations. *Zeitschrift für Psychologie mit Zeitschrift für a ngewandte Psychologie, 195*(2), 129–139. https://psycnet.apa.org/doi/10.2307/1131753.

Loewenstein, J., & Gentner, D. (1998). Relational language facilitates analogy in children. In M.A Gernsbacher & S. J. Derry (Eds). *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 615-620).

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4176-4181. https://doi.org/10.1073/pnas.1814779116.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100(2),* 254. https://psycnet.apa.org/doi/10.1037/0033-295X.100.2.254.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean, & J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, 3111–3119. https://doi.org/10.48550/arXiv.1310.4546.

Morrison, R. G., Doumas, L. A., & Richland, L. E. (2011). A computational account of children's

analogical reasoning: balancing inhibitory control in working memory and relational representation. *Developmental Science*, *14*(3), 516-529. doi: 10.1111/j.1467-7687.2010.00999.x.

Murphy, M. L. (2003). Semantic relations and the lexicon: Antonymy, synonymy, and other paradigms. Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9780511486494.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of a sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology General*, *131*, 5–15. doi: 10.1037/0096-3445.131.1.5

Phillips, C. I., & Pexman, P. M. (2015). When Do Children Understand "Opposite"?. *Journal of Speech, Language, and Hearing Research*, *58*(4), 1233-1244. https://psycnet.apa.org/doi/10.1044/2015_JSLHR-L-14-0222.

Piaget, J., Montangero, J., & Billeter, J. (1977). *La formation des correlates. Recherches sur l'abstraction reflechissante* I, J. Piaget, ed (Paris, Presses Universitaires de France), 115-129.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, *12*(5), 410-430. https://doi.org/10.1002/bs.3830120511.

Quillian, M. R. (1969). The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM*, *12*(8), 459-476. https://doi.org/10.1145/363196.363214.

Rattermann, M.J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: children's performance on a causal-mapping task. *Cognitive Development*, *13*, 453–478. https://psycnet.apa.org/doi/10.1016/S0885-2014(98)90003-X.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*(3), 249-273. https://psycnet.apa.org/doi/10.1016/j.jecp.2006.02.002.

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. https://psycnet.apa.org/doi/10.1016/S0022-5371(73)80056-8.

Sandhofer, C., & Smith, L. B. (2007). Learning adjectives in the real world: How learning nouns impedes learning adjectives. *Language Learning and Development*, *3*(3), 233-267. https://psycnet.apa.org/doi/10.1080/15475440701360465.

Singer-Freeman, K.E. (2005). Analogical reasoning in 2-year-olds: The development of access and relational inference. *Cognitive Development*, *20*, 214-234. doi:10.1016/j.cogdev.2005.04.007.

Smith, C. L. (1979). Children's understanding of natural language hierarchies. *Journal of Experimental Child Psychology*, *27*(3), 437–458. https://doi.org/10.1016/0022-0965(79)90034-1.

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, *24*(1), 99-142. https://psycnet.apa.org/doi/10.1016/0010-0285(92)90004-L.

Smith, L. B., Jones, S. S., & Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. *Developmental Psychology*, *28*(2), 273. https://psycnet.apa.org/doi/10.1037/0012-1649.28.2.273.

Tribushinina, E., van den Bergh, H., Kilani-Schoch, M., Adsu-Koc, A., Dabasinskiene, I.,

Hrzica, G., . . . Dressler, W. U. (2013). The role of explicit contrast in adjective acquisition: A cross-linguistic longitudinal study of adjective production in spontaneous child speech and parental input. *First Language, 33*, 594–616. https://doi.org/10.1177/0142723713503146.

Waxman, S. R. (1991). Convergences between semantic and conceptual organization in the preschool years. *Perspectives on language and thought: Interrelations in development*, 107-145. https://psycnet.apa.org/doi/10.1017/CBO9780511983689.005.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology*, *29*, 257-302. https://psycnet.apa.org/doi/10.1006/cogp.1995.1016.