A Semantics Based Computational Model for Word Learning

by

Ishaan Grover

B.S., Georgia Institute of Technology (2016)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Signature redacted

Author_____

Ishaan Grover
Program in Media Arts and Sciences
August 20, 2018

Signature redacted

Certified by_____

Cynthia Breazeal
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by_____

Tod Machover
Muriel R. Cooper Professor of Music and Media
Academic Head, Program in Media Arts and Sciences

A Semantics Based Computational Model for Word Learning

by

Ishaan Grover

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 20, 2018, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

Studies have shown that children's early literacy skills can impact their ability to achieve academic success, attain higher education and secure employment later in life. However, lack of resources and limited access to educational content causes a "knowledge gap" between children that come from different socio-economic backgrounds. To solve this problem, there has been a recent surge in the development of Intelligent Tutoring Systems (ITS) to provide learning benefits to children. However, before providing new content, an ITS must assess a child's existing knowledge.

Several studies have shown that children learn new words by forming semantic relationships with words they already know. Human tutors often implicitly use semantics to assess a tutee's word knowledge from partial and noisy data. In this thesis, I present a cognitively inspired model that uses word semantics (*semantics-based model*) to make inferences about a child's vocabulary from partial information about their existing vocabulary. Using data from a one-to-one learning intervention between a robotic tutor and 59 children, I show that the proposed semantics-based model outperforms (on average) models that do not use word semantics (*semantics-free models*). A subject level analysis of results reveals that different models perform well for different children, thus motivating the need to combine predictions. To this end, I present two methods to combine predictions from semantics-based and semantics-free models and show that these methods yield better predictions of a child's vocabulary knowledge. Finally, I present an application of the semantics-based model to evaluate if a learning intervention was successful in teaching children new words while enhancing their semantic understanding. More concretely, I show that a personalized word learning intervention with a robotic tutor is better suited to enhance children's vocabulary when compared to a non-personalized intervention. These results motivate the use of semantics-based models to assess children's knowledge and build ITS that maximize children's semantic understanding of words.

Thesis Supervisor: Cynthia Breazeal
Title: Associate Professor of Media Arts and Sciences, Program in Media Arts and Sciences

The following person served as reader for this thesis:

Signature redacted

Dr. Iyad Rahwan

Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences, MIT Media Lab

The following person served as reader for this thesis:

Signature redacted

Dr. Goren Gordon

Head of Curiosity Lab
Tel-Aviv University

Acknowledgements

During the last two years in the Personal Robots Group, I have been able to see research through a lens of camaraderie, passion, humanity and self-discovery. Thank you!

This thesis would not have been possible without the help and support of many people. First and foremost, I would like to thank my advisor Dr. Cynthia Breazeal for being a great advisor whose vision for social robots and AI deeply inspires me. I'm grateful to you for always encouraging me and believing in me. Thank you for providing me with endless opportunities, guidance, support, feedback and the freedom to explore and pursue some rather outlandish ideas over the last two years!

I would also like to thank Dr. Goren Gordon for finding the time to Skype with me, suggesting new ways to approach a problem and providing me with valuable feedback that immensely helped me shape my thesis. Dr. Iyad Rahwan's questions and feedback also helped me in critically thinking about the research questions.

Dr. Hae Won Park, thank you for always being there whenever we needed you! Every time I have randomly popped into your office asking if you had "some time", you've always made yourself available no matter how busy you were. Thank you for always passionately brainstorming with me, helping me "fix bugs" and even meeting me at Clover to discuss ideas whenever I was stuck. You were a constant source of support and encouragement.

Pedro and Nikhita, you've been my greatest friends. I'm lucky to have embarked upon this adventure with you both. I will miss the sleepless nights working on projects, having random yet meaningful conversations on the rooftop, discussing questions about life but most of all, simply being there for each other. Polly, you make the lab come alive with your humor and hugs! Thank you for making us smile even in our most stressful days. We love you! Sam and Sooyeon, you've been the best mentors. You've always been there whenever we've needed help. Thank you for showing us the ropes and being ever so welcoming.

Abhimanyu and Spandan, I met you guys about a year and a half ago and since then you've become my family-away-from-family. We've spent countless hours discussing research, new ideas and pretty much everything under the sun. I will always remember the quizzing nights, breaks while pulling all-nighters, random get togethers, the camaraderie and the friendship! Adam, you've given me a new perspective and helped me learn that there is beauty in exploring without necessarily looking for answers and that some projects are simply meant to bring a smile! Thank you for pushing me out of my comfort zone and exposing me to new experiences. Ishwarya, you were my first friend at MIT! Thank you for being a sounding board and filter for all my weird and insane ideas. Vasundhara, you've been my best friend for over 5 years now. I've called you when I was happy and when I was sad. Thank you for being a constant source of encouragement!

I am also grateful to my first research advisors when I was an undergraduate, Dr. Thad Starner and Dr. Charles Isbell, for accepting me into their labs when I had absolutely no research experience. Dr. Thomas Cederborg, I am grateful to you for initiating me into the world of research, for believing in me, teaching me, mentoring me and giving me the utmost confidence to pursue my ideas without fear of failure.

I would especially like to thank my mom and dad for their unwavering support and faith in me. I have you to thank for everything I am today. No words can express how much your love and support means to me! Thank you for always being there.

This research was supported by NSF IIP-1717362 and NSF IIS-1523118.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"To live in the world without becoming aware of the meaning of the world is like wandering about in a great library without touching the books." - Dan Brown, The lost symbol

Humans are the only species known to have developed the symbol system of language [46]. It gave us the ability to express the myriad conscious experiences we encounter, to transfer information to one another and to *learn* from each other. In a series of experiments, Edward Sapir [52] and Benjamin Whorf [68] showed that language can also influence the way people think and perceive. In one experiment, Whorf showed perceptual differences in identifying the colors "blue" and "green" between native speakers of English and the Tarahumara language. The cause of these perceived differences was attributed to differences in the lexicon of the two languages. Language serves both as a medium for expression as well as perception of *thoughts* and *the meaning of the world around us.*

1.1 Importance of semantics

At the heart of language lies semantics. It explains how various entities are related (or unrelated) to each other. People often use semantic associations to understand new concepts. In the context of language learning, people understand new words by associating them with words they already know. Studies have shown that children, as young as 3 years of age, often form categories among new objects using their shared semantic properties [30]. During

language acquisition, children often develop a model of semantics to map how different words relate to one another. This semantic categorization of words further facilitates the learning of new words [10]. In this thesis, I propose a computational cognitive model that takes into account semantic relationships between words to make predictions about a child's vocabulary knowledge. In the rest of this thesis, I will refer to this model as *semantics-based model* (unless otherwise specified) and models that don't take into account semantic relations between words as *semantics-free models.*

1.2 Motivation

For a child to learn new words and concepts, they must be exposed to new vocabulary, through both written and oral forms of communication. While schools in the United States heavily focus on reading comprehension, in 2013, only 35% of 4th and 38% of 12th grade students tested on the National Assessment of educational progress were found to be proficient in reading [4]. Several studies have shown that children's early literacy skills can impact their ability to achieve academic success, attain higher education and secure employment later in life [25, 40]. Yet, only 40% of eligible pre-schoolers actually attend preschool. The root cause of the problem lies in early childhood education and learning. This can be attributed to family income, lack of education among parents or limited access to a good schooling system. Children that belong to lower socio-economic status (SES) have smaller vocabularies when compared to children from higher SES [25]. This phenomenon has a cascading effect on children from lower SES as they grow older; smaller vocabularies lead to decreased understanding of written and oral communication, which in turn increases this knowledge gap.

With the advent of technology across all socio-economic status groups (eg., smart phones, tablets, etc), there has been a recent surge in the development of Intelligent Tutoring Systems (ITS). These systems are created to provide educational benefits in one-to-one tutoring [66]. For instance, recently the state of Providence launched a program called "Providence Talks" to bridge the word gap between children from low income SES and high income SES [48].

They use a recording device to monitor a child's verbal communication with adults and provide insights into a child's progress in language acquisition. In vocabulary learning tasks, most ITS come with a fixed curriculum (or a set of words) that the system attempts to teach without considering prior knowledge of the tutee. When a human tutor interacts one-to-one with a tutee, s/he attempts to understand existing knowledge of the tutee before teaching a new concept. However, in at-risk communities, it is nearly impossible for a teacher to personalize the curriculum to individual learners. Given the ease-of-deployment, the potential to scale ITS, and recent advancements in the field of AI, there is a significant advantage to building computational cognitive models that personalize to different learners with the goal of maximizing learning.

1.3 Towards building a cognitive model

Whether to understand human cognition or as part of an ITS, the fundamental challenge for a cognitive model is to maintain a belief over possible human behaviors. In the context of language learning for children, a cognitive model must maintain a belief over a child's knowledge. Human tutors often perform this task in the face of sparse, inconsistent, incomplete and noisy data [23].

Consider a hypothetical case where an adult tutor tries to predict whether a child knows the word "Jupiter". If the the tutor identifies that the child knows the word "Earth", does the tutor's belief about the child knowing "Jupiter" become stronger? Now let's assume that the tutor finds out that the child does not know the word "Planet." Does this belief now become weaker? As the tutor talks to the child, the tutor figures out that the child knows "Venus", "Mars" and Neptune"– does the tutor's belief become significantly stronger now? Finally, the tutor identifies that the child knows the word "dog". Does this information change the tutor's belief of whether the child knows "Jupiter"?
Let's deconstruct the thought process of the tutor in the example above:

- The tutor assumed that "Jupiter" was semantically related to "Earth", "Venus", "Mars", "Neptune" and "Planet" but not related to "dog". This was the tutor's

prior knowledge.

- The tutor further assumed that the child must have formed similar semantic relations for the words the child knew.
- The tutor made *observations* about the child's existing knowledge.
- Finally, the tutor updated his/her belief about the child's knowledge of the word "Jupiter" *conditioned on* the observed knowledge from the child.

As humans, we constantly use this *theory of mind* framework to infer beliefs of other people to inform our own. We often make use of semantics in such a framework. For instance, in conversations, we direct conversations to topics where all participants have some background knowledge. In academic papers, we often find the background or literature review section to supply the required semantic information to understand the paper. In fact, this thesis itself is written keeping this fact in mind, and the above example was taken to supply this semantic information!

However, this is not the only framework humans use to make inferences and decisions. For example, the tutor could have assumed that "Jupiter" is not a commonly used word in spoken language, and hence the child is not very likely to know the word. The cognitive process involved in inference, learning and decision making is complex. In literature, several theories for the same have been proposed and each of these theories tries to explain some part of this complex cognitive process. This thesis, however, focuses on the theory of cognition that takes *semantics* into account.

1.4 Research questions

Inspired by this idea of belief formation using semantics, this thesis aims to answer the following research questions:

- **R1:** Given a partial existing vocabulary of a child, can we predict whether s/he would know other semantically related words using a model that uses a theory of mind framework *(semantics-based model)*?

- **R2:** Can we use semantics-free models in conjunction with the semantics-based model to make better predictions about a child's existing vocabulary?
- **R3:** Given an intervention to increase the vocabulary of a child, can we use the semantics-based model to determine how well a child learned new words in the context of other semantically related words?

1.5 Thesis overview

The outline for the rest of the thesis is as follows:

- *Chapter 2*: I highlight key insights on word acquisition among children using semantics from linguistics and developmental psychology.
- *Chapter 3*: I provide a literature review of prior research in modeling semantics and assessing children's vocabulary knowledge.
- *Chapter 4*: I present the experimental design and data collection methodology used to answer the research questions.
- *Chapter 5*: I present details of the computational and mathematical formulation of the semantics-based model as well as semantics-free models. I further provide details on how well each of the models performed on the dataset. I show how the semantics-based and semantics-free models can be combined to make better predictions. Finally, I provide a detailed analysis and discussion of the results presented.
- *Chapter 6*: I present how the semantics-based model can be used to evaluate the success of *personalized* (personalized to the learning level of a child) and *non-personalized* interventions in increasing children's vocabulary.
- *Chapter 7*: I provide a summary of the contributions of this thesis and directions for future work.

Chapter 2

Background

"Everyday language is a part of the human organism and is no less complicated than it" - Ludwig Wittgenstein, Tractatus Logico-Philosophicus

Humans start showing signs of language acquisition as early as infancy [26]. Children as young as four months of age are able to distinguish between their native language and a foreign language based only on auditory cues [34, 38]. By the age of seven months, they are able to identify individual words from a stream of speech [50]. Alongside taking these primary steps to language acquisition, children perform other cognitive processes such as *(i)* parsing new inputs (e.g., words, sentences) to understand underlying concepts [65], and *(ii)* storing the newly acquired information into memory for *retrieval* and *learning new concepts* in the future [29]. What is even more astonishing is that children perform these tasks adeptly in unstructured environments from noisy and multimodal data. As a result, by the age of 6 years, an average child has a vocabulary of approximately 14,000 words [6]. Several theories have been proposed to explain language acquisition in children. Insights from these theories help us understand the complex process of learning, and inform the development of device technologies to help children learn more efficiently.

In this chapter, I give an overview of the different psycholinguistic theories of learning in children. I then give a brief overview of different models of cognition, and the cognitive theory in which I ground the assumptions of the proposed cognitive model.

2.1 Psycholinguistic theories for language acquisition in children

In the early 1950s, many believed in the **behaviorist theory** of language learning that posited language was acquired through imitation and reinforcement [56]. This theory assumed that humans learn language through classical conditioning where verbalization was considered akin to behavior. It was then reinforced by observing the use of language in the person's environment *(stimulus)*. This theory was popularly debunked by another theory that argued that humans did not learn language like other behaviors, instead they were born with innate **pre-existing cognitive mechanisms** which allowed for the acquisition of language [8, 44]. Another cognitive theory of learning argues that children learn words by **forming associations** with objects they observe [58]. For instance, if a child saw a "house" and heard someone refer to it as a "house", they would associate the word with the actual concept of house. Research has shown that even if a child looks at the picture of a house instead of looking at an actual house, they are able to associate the word with the concept instead of the picture [45].

However, this theory doesn't explain how children form associations in the presence of noisy stimulus. For example, how does the child know that the person was referring to the house and not the "door" to the house. This problem of uncertainty of word to concept mapping because of noisy stimulus is known as referential uncertainty [47]. It is argued that in such instances, children use **non-verbal behavior** to identify the referent of a word [7]. The theory of **cross-situational learning** deals with the problem of referential uncertainty in a different manner. It posits that even though there might be multiple word-concept mappings possible in a given situation (or sentence), the word-concept mapping for a given word remains constant across different uses of the word. Thus, children are able to learn the concept corresponding to a given word by keeping track of the mapping that remains constant across different situations [57].

On the other hand, the theory for **incidental learning** argues that language learners often acquire vocabulary through reading, listening and conversational activities [27].

During such activities, a learner is not actively trying to learn new words, but instead *incidentally* "picks up" new words since the listener's attention is on the meaning rather than the grammatical structure of language [27]. In first language learners, this kind of learning takes place for the majority of their lives as children are exposed to their native language in homes, schools and other social contexts. Research has further shown that during incidental vocabulary acquisition, a learner's vocabulary growth is more from listening than reading [61]. Moreover, incidental vocabulary acquisition has a direct link with frequency of exposure, i.e, the more a child is exposed to a given word, the higher are the chances of the child committing the word to memory [63]. As a corollary, one may argue that a child is more likely to know a word that has a high frequency of occurring in everyday use of language. For example, it is more likely for a child to know the word "cat" rather than "feline".

Apart from frequency of word usage, **phonological similarity** of words also plays a role in retention, i.e., using phonologically similar words can be used to aid retention [54]. However, a phonologic similarity effect is often observed in children which shows that recall of phonologically similar words is slower than that of phonologically dissimilar words [11]. Thus, children are often taught rhyming words and poems with a rhyme-scheme to increase their phonological awareness; higher sensitivity to rhyme positively affects awareness of phonemes and reading skills [1]. This strategy in turn increases exposure to such words which might induce incidental learning among children. The phonologic similarity effect is reversed when additional semantic information is provided [12]. This motivates the fact that there might exist cognitive mechanisms that aid in learning, storage and retrieval of semantic information.

Research has shown that during comprehension process, people retrieve previously learned vocabulary words in order to make new conceptual associations [29]. Further, a study showed that when people paid attention to sound, appearance and semantics of a word in three different conditions, their retrieval time for a given word was least when they paid attention to semantics [28]. Wolf et al. used semantic representations in the Ravo-o intervention claiming that it played a significant role in word recognition and comprehension [69]. This

intervention showed a significant increase in vocabulary and text-understanding when compared with children who were offered other forms of or no interventions. More recently, researchers conducted a study that showed that when children learned vocabulary along with semantics, they learned more words and their learning pace was faster [14]. Another study compared children who had language comprehension difficulties with a control group of children. The results of the study showed that despite having similar phonological abilities, children with comprehension difficulties had problems reading words that were often read with support from semantics [37]. This shows that children with typically developing language comprehension skills have a semantic understanding of how words relate to one another. These studies motivate the fact that *(i)* children learn better when taught words that are semantically related, and *(ii)* children form a cognitive semantic representation for retrieval from memory.

Thus, there are several proposed theories that try to explain how children acquire language. No one theory is correct or wrong by itself. Each theory tries to explain some part of this huge puzzle of cognitive mechanisms that allow acquisition, storage and retrieval of language and concepts.

2.1.1 Representation of semantic knowledge

For three decades, cognitive science has emphasized the importance of representing knowledge in the form of data structures. The **schemata theory** presents a knowledge representation quite similar to that of semantic networks. It claims that this representation is the basic building block of cognition. In fact, in 1980, it was suggested that language acquisition follows a cognitive model based on schemata theory and that vocabulary development should take place simultaneously with background knowledge [49]. Thus, it is important to represent knowledge semantically as it helps understand how new information is processed, added and retrieved.

A common representation of semantic knowledge is a graph structure called a semantic network. The nodes in the graph represent concepts and the edges represent the semantic

relationship between those concepts. In the context of language learning, nodes can be seen as analogous to words and edges can be seen as analogous to relationships between those words. We choose words instead of smaller units, say morphemes, because we use words in our daily life, teach words intentionally and think about words when relating concepts. For example, one would relate "table" and "chair" but not "un", "break" and "able".

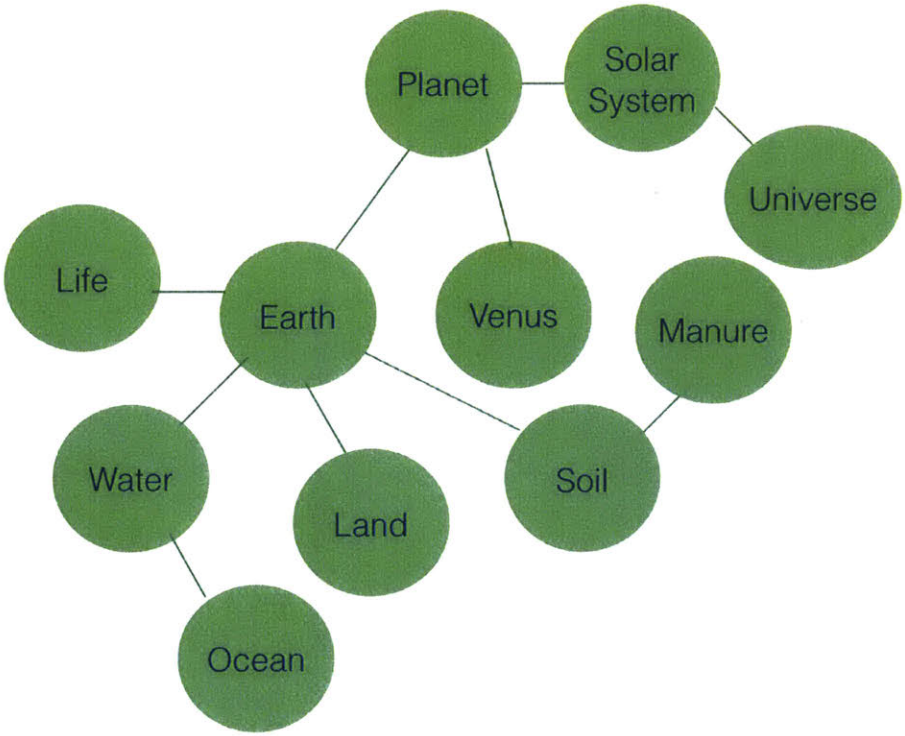

Figure 2-1: Example of semantic network of words

2.2 Models of cognition

A cognitive model tries to explain some part of human decision making or problem solving to *simulate* or *predict* human behavior. These models often make some assumptions that are grounded in theories of cognition. Since, no theory can completely describe how the hu-

man brain works, the capabilities of these models are constrained by the set of assumptions they make. A computational cognitive model is cast at Marr (1982) implementation level of analysis of information processing systems [23]. Several models for cognition have been proposed to model how humans make inferences in intelligent tasks such as decision making, generalization and concept learning among many others [33, 62, 51].

The computational-representation understanding of mind (CRUM) hypothesizes that there exists representations in our minds over which thinking is performed, that these representations can be seen analogous to data structures, and thinking may be seen as analogous to algorithms [64]. Under this taxonomy, the model proposed in this thesis is based on the semantic theory of vocabulary acquisition among children. More specifically, the model tries to make inferences, akin to those made by a human tutor, that assumes the semantic model of vocabulary acquisition for its tutee. Thus, the fundamental assumption made by this model in order to make inferences is the following:

H1: *Children learn words by forming semantic relations with existing words that they know. Thus, if it is observed that a child knows a word, it is likely with some probability that the child knows words that are semantically related to the given word. On the other hand, if it is observed that the child does not know a given word, there is some probability that the child does not know words that are semantically related to the given word.*

Chapter 3

Related work

The model proposed in this thesis takes inspiration from and builds upon three main research areas, namely: *(i) cognitive models for vocabulary acquisition, (ii) semantic categorization and (iii) knowledge prediction.* In this chapter, I provide an overview of existing work in each of these research areas and discuss how this thesis contributes to the existing body of work. The first research question is re-stated here to situate the proposed model within the context of existing research.

R1: Given partial existing vocabulary of a child, can we we predict whether a child would know other semantically related words using a model that uses a theory of mind framework? *(semantics-based model)*?

3.1 Cognitive models for vocabulary acquisition

Current computational cognitive models aim to model some aspects of vocabulary acquisition based on cognitive theories of word acquisition (as discussed in Chapter 2). For instance, Siskind (1996) proposed a model that used the principles of cross-situational learning to learn word-to-concept mappings. The model made the assumption that when children hear new utterances, they use existing knowledge to constrain the hypothesis space for each word-to-concept mapping [55]. Frank et al. [17] attempted to solve the problem of referential uncertainty by performing inference to find speaker's intentions and word-to-concept

mappings simultaneously. Yu et al. [70] took a different approach by incorporating social cues like joint attention into a unified statistical learning framework for cross-situational observations. These models, however, did not take into account semantic relationships between words. More recently, a model to build and incrementally grow a semantic network from utterance data was proposed to capture semantic relationships between words [39]. A common feature of these models is to learn word-to-concept mappings from data just as a child would.

However, in our case, we are trying to *predict* words that a child might know based on knowledge of a given set of words. The learning of words has already happened and we do not have access to the data stream that the child was exposed to when they learned the words that exist in their current vocabulary. Thus, our goal is to build a cognitive model for a person trying to predict whether the child knows some given target word, based on the cognitive theory that children learn words by forming relations with semantically related words that they already know.

3.2 Semantic categorization

For over two decades, building computational models for the semantic categorization of words and objects has been an active area of research. Early work in building a computational cognitive model involved a bayesian learning framework to incrementally update the posterior probability of an object belonging to a category given some partition of the feature set [2]. In the context of semantic categorization of words, perhaps the most popular model is Latent Semantic Analysis (LSA) [32]. It performs dimensionality reduction on a word frequency-document matrix (using SVD) to find a point for each word in an n-dimensional space. The distance between two points gives a measure of semantic similarity of the words. Subsequent research in *topic models* suggested using generative models to first sample a topic and then sample a word from the topic [22]. Solving the inference problem would then give a vector representation for each word. With recent technological advancements that provide high computational power, unsupervised deep learning methods to learn semantic vector spaces have gained popularity [43, 35]. Similar to LSA, these models output a vector

representation for each word, and the distance between two vectors encodes the semantic similarity between two words. The model proposed in this thesis requires building of a semantic network and it relies on GloVe embeddings to find semantically similar neighbors. A detailed description of the model is provided in Chapter 5.

3.3 Knowledge prediction

Within the ITS literature, researchers have proposed models for assessing children's reading skills and pronunciation. For instance, Gordon et al. [19] presented a bayesian active learning-based model to predict the probability of a child's ability to *read* a word. They used bayes's rule to compute the probability of being able to read a word and a phonetic distance based heuristic to estimate the conditional probability of a child knowing how read a word w_1 given that they know how to read some word w_2. Spaulding et al. [60] showed that incorporating affect information (smile and engagement) into the BKT model outperforms traditional models for predicting a child's reading skill level. Researchers have also used gaussian process regression to model children's *pronunciation skills* using a covariance function that is a weighted sum of semantic and phonemic similarity [59].

However, there has been little work in predicting *word understanding* through semantics. Recently, researchers tried to model word learning using the Osgood semantic scale and word2vec embeddings [36]. The Osgood scale maps every word to the same low dimensional feature space (10 features). For instance, some of the features are, "on a scale of 1-10, how good/bad is the given word?", or "on a scale of 1-10, how active/passive is a given word?". They then use mixed effect logistic regression to predict short and long term word acquisition in children. However, using the Osgood scale requires taking a survey to get a Likert rating on each dimension for every word for which we wish to make a prediction. This method becomes infeasible when there are a large number of words in the corpus. For the word2vec embeddings, the features used in the model are dependent on the design of the study. This requires for each target word, testing the participant's knowledge about semantically related words when used in the same sentence as the target word. This motivates the development of a model that is based on semantic word similarities and independent of the experimental

protocol.

In chapter 5, I detail a **model of cognition** for an observer making **predictions** about a child's vocabulary knowledge using their own beliefs about the child's learning of vocabulary using **semantic word similarities**.

Chapter 4

Experimental design

In this chapter, I give details of the experiment and the data collected. Finally, I re-state the research questions more specifically in terms of the data collected.

4.1 Overview

The experiment consisted of three parts, namely:

- Part 1: Pre-test
- Part 2: Intervention
- Part 3: Post-test

At the beginning of the experiment, subjects were asked to take a small test as a means to measure their current vocabulary level and to sample words from their existing vocabulary. After the pre-test, subjects were divided into three conditions: *(i)Baseline*, *(ii) Personalized*, *(iii)Non-Personalized* such that the vocabulary level of subjects was counterbalanced to be approximately the same across conditions. Each condition specified the kind of teaching intervention subjects would go through. At the end of the intervention, subjects from all conditions were asked to take a post-test to again measure their vocabulary level and sample new words from their existing vocabulary. A flow chart of the experiment is provided in Figure 4-1. Details about each part of the experiment are provided in the subsequent sections of this chapter.

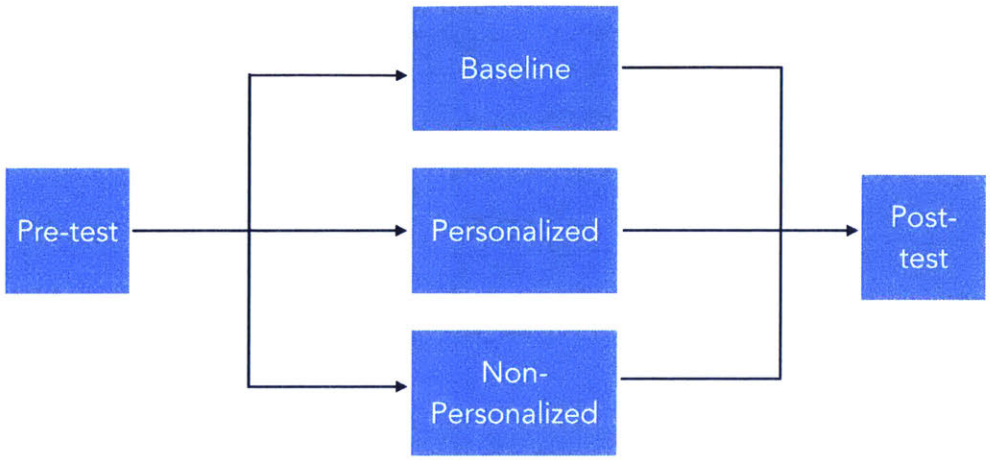

Figure 4-1: Flow chart of experiment design

4.2 Participants

We first recruited 73 English language learners (ELL) and bilinguals from local public schools in the Greater Boston area for a three-month long experiment. Afterwards, six children who earned significantly higher vocabulary scores in the pre-test compared to other children ($p = 0.013$) were excluded from the study. Given the longitudinal nature of the deployment, another 8 children were unable to complete the experiment to take the final post-test. These participants were also excluded from the study. In total, 59 children between the ages of 4–6 (female 54.2%, age $\mu = 5.33$, $\sigma = 0.63$) from 12 classrooms participated in the study. The number of children in each condition is given in table 4.1.

Table 4.1: Number of subjects per condition

Condition	Number of subjects
Baseline	17
Personalized	22
Non-personalized	20

4.3 Pre-Test and Post-Test

It is hard to measure a child's current vocabulary level since the set of words a child could know is very large, and asking a child the meaning of a large set of words is infeasible. The goal of the pre-test was twofold: *(i) measure the vocabulary level of the child (to divide children into different conditions)* and *(ii) sample words from a child's vocabulary.* Each child's linguistic level was assessed using a clinically evaluated vocabulary test and a personal narratives collection protocol. We used the Peabody Picture Vocabulary Test (PPVT) [15] which is one of the most widely used tests to measure a child's vocabulary level. The format of the test involves asking a child to select (or point to) one of four possible pictures that are related to the target word being assessed. The words presented in the test are such that they align with the vocabulary level of the child. The test ends when the words presented are above the vocabulary level of the child, and the child is constantly unable to answer the questions. Here, vocabulary level was determined by conducting large surveys (> 2000 people) and assessing how different populations across different geographic location, age groups, etc. performed on the answering questions about different words in the test.

After repeating this process for multiple words in the test, we are able to obtain a dataset containing a sample of words that a child knows and doesn't know as well as a measure of the child's vocabulary. Index of Productive Syntax (IPSyn) [53] evaluates the grammatical complexity of spontaneous language samples. Additionally, we evaluated children's syntax complexity from collected personal narrative samples using IPSyn's noun phrase, verb phrase, and sentence structure scales. The personal narratives collection followed Westerveld and Gillon's language sampling protocol [67]. Together, using the IPSyn scores and PPVT scores, children were divided into three groups such that no group was skewed in terms of the vocabulary level of children to start.

The goal of the post-test was to re-sample words from a child's vocabulary after the intervention. Thus, a similar procedure of using PPVT after the intervention was performed. The set of target words assessed during the post-test were completely different from those used in the pre-test. One word was common between the two tests, and a design choice of including it only in the pre-test was made.

In addition to words assessed in PPVT, we also hand-picked a set of 25 target words and tested children's knowledge of these words in the pre-test. These target words were then also carefully embedded in the stories told to the child. Finally, during post-test, we again assessed children's knowledge of these words to check if they learned these words as a result of the intervention. Assessment of these words was done using the same format as PPVT.

4.4 Intervention

4.4.1 Robot Platform

Tega is an appealing, expressive, child friendly robot, designed for long-term deployment in various educational settings such as children's homes, schools, and therapeutic centers [42, 20]. It is 11 inches tall with a cylindrical body and is brightly colored with a plush exterior. It has five degrees of freedom to perform a wide range of expressive movements: head tilt up/down, waist tilt left/right, waist lean forward/back, body extension up/down, and body twist left/right. An Android smartphone mounted in the head is used to graphically display the robot's animated face as well as perform computational tasks such as sensor processing, data collection, WiFi communications, decision making, and motor control. The robot's electronic design extends the smartphone's ability with an additional high-definition camera with a wide field of view.

Tega is designed to act as a peer-like social robot. A peer is member of a group of people comprised of similar ages and that can influence each other's language, behavior, and beliefs through interaction. Tega's peer attributes include child-like high pitched voice, exaggerated body and facial expressions, and backchanneling response generation to signal engagement while attending to the user [41]. With its peer attributes, Tega is able to engage in taking turns telling stories with a child while encouraging each other.

4.4.2 Interaction with robot

Depending on each school's curriculum, children participated in 6–8 sessions of storytelling with the Tega robot. Each session with Tega started with a short greeting that consisted of relational dialogues including personal, social, and emotional experiences. The session then

progressed into Tega and the child taking turns telling stories to each other. Tega always took the first turn picking and telling a story to the child, followed by the child retelling the story back to Tega using the same storybook (only containing graphics). Children in the baseline condition did not have any interaction with Tega and hence did not partake in any story-telling sessions. In the personalized condition, Tega narrated stories such that the stories were only slightly above the child's linguistic abilities and maximized engagement with the robot using a personalized reinforcement learning algorithm (building a personalized policy for each child) to enhance learning outcomes. In the non-personalized condition, Tega narrated stories that were not personalized to each child and did not rely on any learning algorithm to maximize learning. In both, the personalized and non-personalized interaction, the robot made use of a tablet to display pictures to enhance the storytelling experience of the child.

4.4.3 Story Corpus

In total, 81 storybooks with an illustration on every page were collected. Together, they spanned the spectrum of lexical and syntactic complexity. Most of these storybooks were selected from Tufts University's Reading and Language Research Center, but all are commercially available. Every book in the corpus was evaluated in terms of lexical and syntactic complexity using the automatic assessment tools. Only the graphics from the storybooks were presented on a tablet screen during the robot's and child's storytelling.

Four books were referenced from Collins' article on preschool English Language Learner (ELL)'s vocabulary acquisition from reading storybooks [9]. Researchers at Tuft's Reading and Language Research Center authored four varying levels of these four storybooks. The personalization policy guided the selection of the story level for children in the personalized condition. Children in the non-personalized condition heard stories of alternating levels in alternating weeks. That is, they heard level 1 of the storybook on week 2, level 2 on week 4, level 3 on week 6, and level 4 on week 8, level 4 being the most difficult in terms of lexical and syntactic complexity. In weeks 1, 3, 5 and 7, they heard stories that were not evaluated for level but roughly corresponded to very easy "Level 1" stories.

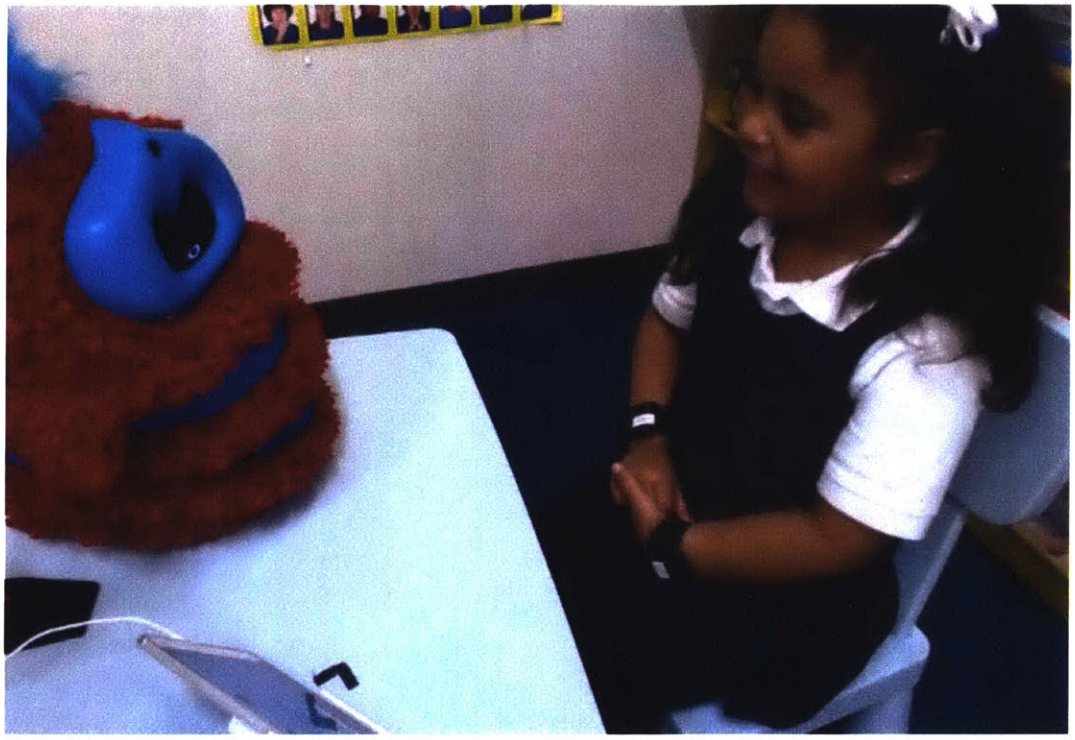

Figure 4-2: Child-robot storytelling interaction

4.5 Restating research questions

Given the experimental design, I now restate the research questions (posed in Chapter 1) more specifically in terms of the data collected from the experiment:

- **R1:** Given a child's knowledge about words assessed in the pre-test, can we we predict whether a child would know other *semantically related* words as assessed in the post-test using a model that uses a theory of mind framework *(semantics-based model)*?
- **R2:** Can we use semantics-free models in conjunction with the semantics-based model to make better predictions about words assessed in the post-test?
- **R3:** Given data about stories that the robot narrated and data from a child's story retell, can we use the sematics-based model to determine how well a child learned new words in the context of other semantically related words?

I define how words were determined to be *semantically related* in the next chapter.

Chapter 5

Semantics-based and Semantics-free models

In this chapter, I address the first two research questions: *(i) R1: Given a child's knowledge about words assessed in the pre-test, can we we predict whether a child would know other semantically related words as assessed in the post-test using a model that uses a theory of mind framework?* and *(ii) R2: Can we use semantics-free models in conjunction with semantics-based models to make better predictions about words assessed in the post-test?*. I first provide an overview of the dataset used to evaluate the model. I then introduce background information about undirected graphical models and cast the first research question as an equivalent computational inference problem on undirected graphical models (semantics-based model). I further explain the assumptions made by the model to make computations tractable. Then I describe other semantics-free models that may be used to answer the research questions. I then provide details on how predictions from semantics-free and semantics-based models may be combined to evaluate if better predictions can be made about a child's vocabulary. Finally, I describe the evaluation methodology, present the results and provide an analysis of the results in the context of the posed research questions.

5.1 Dataset and terminology

The data collected from the experiment is summarized as follows:

- Pre-test: Set of target words W_{pre} (within the expected vocabulary level of a child) and information about whether or not they know those words before a teaching intervention.

- Intervention: Set of words $W_{robot,story}$ that the robot used in telling stories to a child and set of words $W_{child,story}$ that the child used to narrate the story back to the robot.

- Post-test: Set of target words W_{post} (within the expected vocabulary level of a child) and information about whether or not they know those words after a teaching intervention. Moreover, $W_{post} \cap W_{pre} = \emptyset$.

- Common words: Set of words W_{common} that refers to words that were assessed during pre-test, used in the intervention, and again assessed in the post-test.

5.1.1 Measuring semantic similarity between words

In order to answer R1, the term *semantically similar* must be properly defined and quantified. Intuitively, we know that the words "dog" and "cat" are semantically similar and the terms "shoe" and "computer" are not semantically similar. Thus, in order to formalize this notion, we need a function s(w_1, w_2) which quantifies how semantically similar two words are. Firth famously stated that "a word is characterized by the company it keeps" [16]. The distributional hypothesis in linguistics is based on this concept and states that words used in similar contexts tend to be semantically similar [24]. Most word embeddings trained using GloVe, word2vec or skip-gram try to capture this notion using co-occurrence between words to find word vectors in an n-dimensional space. A common measure of semantic distance between words is to take the cosine distance between the word embeddings of two words [59]. Here, I define semantic distance between two words as the cosine distance between their pre-trained common crawl GloVe word vectors (300 dimensional) [43]. Two words with vector representations v_1 and v_2 are said to be semantically similar if $cos(v_1, v_2) \geq \epsilon$

(through empirical investigation of semantic distances between various words, ϵ was found to be 0.6).

5.1.2 Building semantic network

A human tutor often has some prior knowledge about how words are semantically related to each other and a rough estimate of the kinds of words a child might know. For example, even though a human tutor might know the semantically related terms "air", "pressure" and "Gay-Lussac's law", the tutor doesn't assume that the child would know the pressure law when making inferences about other words that a child might know. Thus, the tutor implicitly tries to make an *educated guess* to reduce the kinds of words they would consider when making a prediction. The tutor however does not know for sure which of the words the child actually knows. The closer this educated guess is to a child's actual knowledge, the better inferences the tutor is able to make.

As noted in Chapter 2, a common representation of semantic knowledge is semantic networks. Hence, to represent this belief about a child's knowledge computationally, we build a semantic network using the first thousand words w_{list} that are often used in children's textbooks [18] along with words in W_{pre} and W_{post}. This may be seen as a way to add domain knowledge to the system. Let this graph be called $G_{semantics} = (V, E)$ such that $V = W_{pre} \cup W_{post} \cup w_{list}$. Set of edges E of the graph are computed using pairwise comparisons between nodes to check for semantic similarity of words, i.e., nodes v_1 and v_2 ($v_1 \in V$ and $v_2 \in V$) are connected by edge e_{12} if and only if v_1 and v_2 are semantically similar. Building this graph runs in $O(|V|^2)$ which was computationally acceptable in our case ($|V| \approx 1400$) . It is important to note that $G_{semantics}$ is only a partial representation of knowledge and does not represent any child's "actual" semantic network but instead only aids in reducing the hypothesis space of the kinds of words a child has some non-zero probability of knowing.

5.1.3 Selection of post-test words

The semantic representation of words using $G_{semantics}$ allows us to select words, $W_{post,selected} \subseteq W_{post}$ that are semantically related to words in W_{pre}. One strategy is to only select words from W_{post} that have a direct edge to a word in W_{pre}. However, this selection of words

only allows us to make predictions about words that are *strongly* related to W_{pre}. Another strategy is to perform breadth first search (BFS) to find the shortest path between a node in W_{post} and nodes in W_{pre} and only consider nodes in W_{post} that have a path length $< k$ ($k \in N$) where k is some constant. This selection of words allows us to make predictions about words that are *somewhat related* to words in W_{pre} through $k+1$ nodes. However, we make a design choice to make a predictions about all words in W_{post} that have some path to a node in W_{pre} *regardless* of path length. This strategy only eliminates words from W_{post} that are in no way associated with any node in W_{pre} through any number of connections. Thus, we make inferences about nodes even if they are *weakly* connected to nodes for which true knowledge of a child is known.

It is important to keep the difference between *semantically similar* and *semantically related* in mind. We define two words w_1 and w_2 to be semantically related if there exists a path between them in $G_{semantics}$. They are said to be semantically similar if the cosine distance between their word vector representation $\geq \epsilon$. Thus, a node is semantically related to another node through a path of semantically similar nodes.

To summarize the terminology used, we defined the following terms in this section:

- W_{pre}: set of words in evaluated in pre-test for each child.
- $W_{robot,story}$ and $W_{child,story}$: set of words the robot and the child used in their respective stories.
- $W_{post,selected}$: set of words that are semantically related to words in W_{pre} for each child
- $G_{semantics}$: semantic network representing words and their relations

5.2 Computational formalism

In this section, I show how research problem R1 can be answered by formulating it as an inference problem on undirected graphical models (UGMs). I then explicitly state the assumptions made by the model in making such inferences.

5.2.1 Preliminaries - Markov Random Field

A graphical model is a graphical representation of random variables that encodes assumptions of conditional independence between nodes. This family of models is often used to find marginal distributions and conditional marginal distributions when the joint distribution has special properties of conditional independence [31]. A Markov Random Field (MRF) is an undirected graphical model defined by graph $G_{mrf} = (V_{mrf}, E_{mrf})$ and a set of random variables X where vertices $V_{mrf} = \{v_1, v_2, v_3...v_n\}$ correspond to random variables $X = \{X_1, X_2, X_3...X_n\}$. The edges E_{mrf} are used to define the markov blanket of a node. The markov blanket of a node v is defined as the immediate neighbors of v in $G_{semantics}$. Let $N(v)$ give the neighbors of node v. Then, for a Markov Random Field, we have the following properties:

*(i):*Two non-adjacent variables are independent given all other nodes:

$$X_u \perp\!\!\!\perp X_v \mid X_{V\setminus\{u,v\}} \tag{5.1}$$

*(ii):*A variable is conditionally independent of all variables given its neighbors.

$$X_i \perp\!\!\!\perp X_{V\setminus \mathrm{N}(i)} \mid X_{\mathrm{N}(v)} \tag{5.2}$$

An edge e_{ij} $(i, j \in [1, |X|], i \in N$ and $j \in N)$ between nodes X_i and X_j captures the notion of conditional dependence between nodes. This dependence is numerically represented by a potential function $\phi(\mathbf{x})$ which may be defined for a pair of nodes or a clique (c) in the graph. When it is defined for pairs of nodes, the MRF is called a pairwise MRF. Given the above properties for an MRF, we have the joint distribution:

$$p(X_1, X_2...X_n) = \frac{1}{Z}\prod_{C}\phi_c(\mathbf{x}_c) \tag{5.3}$$

where,

- C is the set of all maximal cliques in the graph

- $\phi_C(x_c)$ is the potential function associated with clique c
- $\phi_c(\mathbf{x}_c) \geq 0$
- Z is the normalizing constant or the partition function

$$Z = \sum_{\mathbf{x}} \prod_{C} \phi_c(\mathbf{x}_c) \tag{5.4}$$

5.2.2 Model formalism - semantic network to MRF

The graph structure of a semantic network is similar to that of a Markov Random Field. Here, I show how a semantic network can be cast as probabilistic graphical model by adding the constraints of an MRF. We create a graph $G_{mrf} = (V_{mrf}, E_{mrf})$ from semantic network $G_{semanics} = (V_{semantics}, E_{semantics})$ such that $|V_{mrf}| = |V_{semantics}|$ and $|E_{mrf}| = |E_{semantics}|$. Every node $V_{i,semantcs}$ in $G_{semantics}$ is mapped to a random variable $X_{i,mrf}$ in G_{mrf} with support {0, 1} (0: does not know the word, 1: knows the word) which represents the distribution of whether or not a child knows the corresponding word in $G_{semantics}$. Two nodes $V_{i,mrf}$ and $V_{j,mrf}$ are connected in G_{mrf} if and only if their corresponding nodes are connected in $G_{semantics}$.

Potential function: Hammersely-Clifford theorem states that if a joint probability distribution that can be represented by a Markov Random Field as a product of clique potentials, then it can also be represented by a corresponding Gibbs Random Field. In other words, the product of potential functions can be written as a product of exponentials. Given this theorem, we have:

$$\phi_C(\mathbf{x}_c) = e^{-E(\mathbf{x}_c)} \tag{5.5}$$

Where $\phi_C(\mathbf{x}_c) > 0$ and $E(\mathbf{x}_c)$ is the energy function associated with the clique $\mathbf{x}_c$. This finally gives us:

$$p(X_1, X_2...X_n) = \frac{1}{Z}\prod_C e^{-E(\mathbf{x}_c)} \tag{5.6}$$

and

$$Z = \sum_{\mathbf{x}}\prod_C e^{-E(\mathbf{x}_c)} \tag{5.7}$$

Thus, the lower the energy of a clique, the higher will be its potential and thus higher will be the probability.

Since we only have a measure of semantic similarity between two words, we use pairwise potential functions to capture the notion of how to words are semantically similar to each other. More specifically, we try to capture the following assumption stated in Chapter 2:

H1: Children learn words by forming semantic relations with existing words that they know. Thus, if it is observed that a child knows a word, it is likely with some probability that the child knows words that are semantically related to the given word. On the other hand, if it is observed that the child does not know a given word, there is some probability that the child does not know words that are semantically related to the given word.

In order to represent this assumption in the model, potential functions must further be associative in nature. That is, they should favor neighboring nodes to have the same label (knowing the word or not knowing the word) giving rise to a model that is both, pairwise and associative in nature.

Given the form of the potential function, we define the energy function as follows: if X_i corresponds to the word w_i and X_j corresponds to the word w_j in graph $G_{semantics}$, and $s(w_j, w_j)$ gives the semantic similarity[1] between words w_i and w_j. We define the energy of

[1]When computing semantic similarity, we use the lemmatized form of words. This is done so that different forms of words don't affect their semantic similarity. For example, semantic distance between "cats" and "dogs" should be the same as that between "cat" and "dog".

neighboring nodes having the same label as:

$$E(X_i, X_j) = 1 - s(w_j, w_j) \tag{5.8}$$

and energy of neighboring nodes having a different label as:

$$E(X_i, X_j) = s(w_j, w_j) \tag{5.9}$$

Thus, the higher the semantic similarity between two nodes, the lower will be the energy associated with the pair of words. Alternatively, the lower the semantic similarity between words, the higher will be the associated energy. This results in the final pairwise potential function:

$$\phi(X_i, X_j) = \begin{bmatrix} e^{-(1-s(w_i,w_j))} & e^{-s(w_i,w_j)} \\ e^{-s(w_i,w_j)} & e^{-(1-s(w_i,w_j))} \end{bmatrix} \tag{5.10}$$

This form of the potential function ensures that in the absence of any observations, marginals for all nodes is uniform after LBP is performed. Thus, the model starts with an initial prior of assuming an equal probability of knowing or not knowing any word (0.50).

5.2.3 Inference - MRF to factor graph

Given a probabilistic graphical model (representing the joint distribution in the form of conditional independence relations) and observations of some nodes, we find conditional marginals of any of the remaining nodes using the process of *inference*. The graph structure of an MRF may or may not contain cycles. Depending on the structure, an appropriate algorithm is used for inference. Given the nature of the problem, it is likely that G_{mrf} would contain cycles. For instance, three words could all be semantically similar to one another forming a cycle. For a graph containing cycles, exact inference is computationally intractable. In such cases, marginal distributions of nodes can only be approximated. A common algorithm to perform approximate inference on graphs containing cycles is called Loopy Belief Propagation (LBP). In this thesis, I use the sum-product variant of LBP. Since

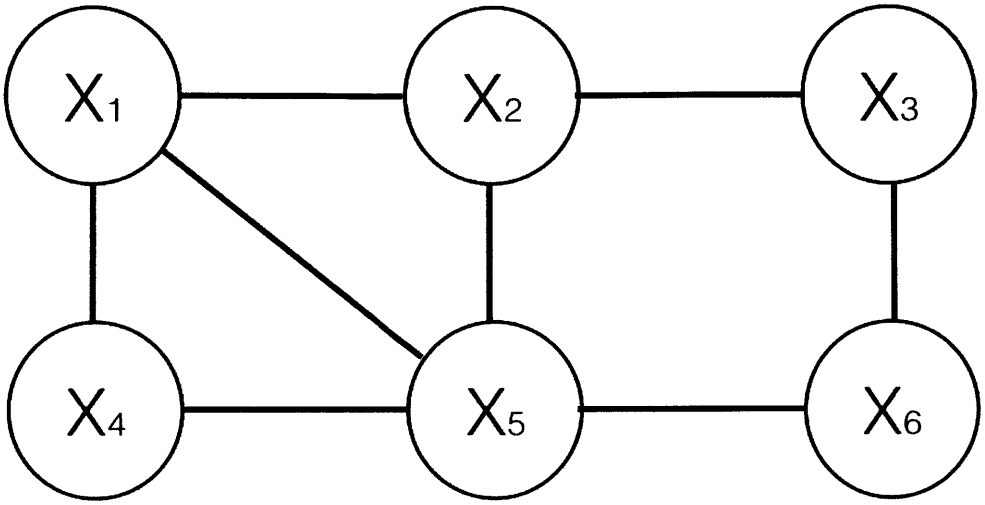

Figure 5-1: Example of a markov random field

Loopy Belief Propagation is computationally expensive, I reduce the size of the semantic network by only considering nodes that lie along the shortest path from each word in $W_{post,selected}$ to W_{pre}.

An MRF is often represented as a factor graph to perform LBP to compute conditional marginals of nodes. Thus, from graph G_{mrf}, we construct factor graph $G_{fg} = (V_{fg}, F, E_{fg})$ in the following way: $V_{fg} = V_{mrf}$. Set of factor nodes F represent the pairwise potentials between connected nodes. That is F_{ij} represents the pairwise potential ϕ_{ij} in G_{mrf} between nodes $V_{i,mrf}$ and $V_{j,mrf}$. E_{fg} is a set of edges such that if an edge exists between any pair of nodes $V_{i,mrf}$ and $V_{j,mrf}$, there exists an edge between $V_{i,fg}$ - F_{ij} and $V_{j,fg}$ - F_{ij} in G_{fg}. Thus, $|F| = |E_{mrf}|$ and $|E_{fg}| = 2|E_{mrf}|$. An example of a markov random random field is given in figure 5-1 and its equivalent factor graph representation is given in figure 5-2.

R1 as an inference problem: Each child's pre-test provides observations of the nodes that a child knows and doesn't know. The goal of the model is then to use pre-test words as observations to find the conditional marginal probabilities of words in the post-test

$W_{post,selected}$ using LBP on G_{fg}. Since each child has a different set of observations, the model finds the marginal probability of knowing words in $W_{post,selected}$ for each child using these different sets of observations.

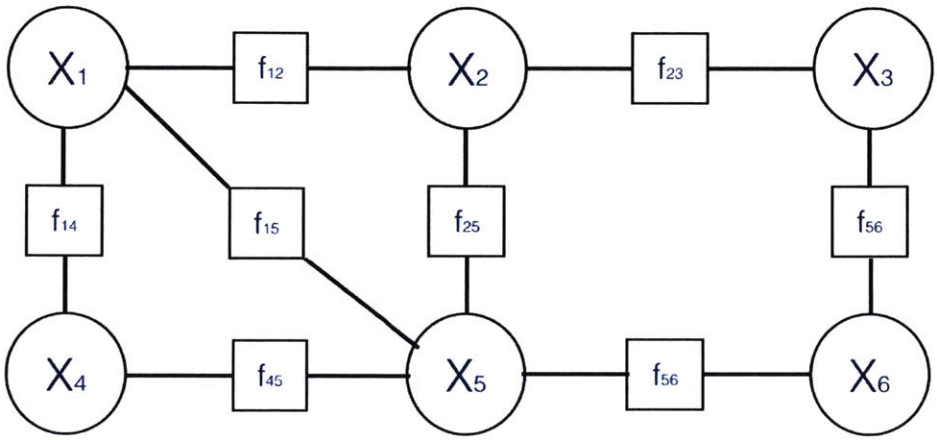

Figure 5-2: Example of factor graph representation

5.2.4 Intuitive explanation of model and assumptions

The associative nature of potential functions probabilistically encodes the intuition that if a child knows a given word, there is some probability that they would know the neighboring nodes (and vice-versa). The model makes a strong assumption that a word is conditionally independent of all other words given its neighbors. However, since every node makes the same assumption, this assumption allows the model to *propagate* belief throughout the network. Intuitively, when we observe that a child knows a given word, the model propagates this positive belief throughout the graph because of associative pairwise potentials. Similarly, when we observe that a child doesn't know a word, a negative belief is propagated through the network. Moreover, the higher the semantic similarity of a node with its neighboring node, the higher will be the effect on it. This effect becomes less pronounced as the path length increases. Hence, the model is able to compute marginals for any arbitrary configuration of observed nodes. It is important to note that the model only uses

the words in w_{list} as nodes that would exist in a child's knowledge base. While this is a reasonable assumption, it is only a weak approximation of the child's true knowledge and more information about the child's vocabulary knowledge would yield better representations, thereby improving prediction performance. In this way, using a theory of mind framework, the model forms a *prior* about a child's existing semantic network and is able to perform inference based on semantic relations between words and *observations* from a child much like the human tutor discussed in Chapter 1.

5.3 Semantics-based baseline models

The proposed graphical model (MRF) has two key components: *(i)* A graph representing semantic relationships which encodes *priors* and *(ii)* the ability to perform inference. In this section I discuss models that use semantics but lack some of the features of the model proposed in the previous section. I use these models or algorithms as baselines to later compare the performance of the graphical model.

5.3.1 GloVe nearest neighbor

In this section, I discuss a model that uses word semantics but does not use the semantic network $G_{semantics}$ or inference algorithms to make predictions. Thus, to predict whether a child knows a given word or not, we find the word in W_{pre} that has the highest measure of semantic similarity (word with the highest cosine distance between GloVe vector representations) with the given word and assign the label corresponding to the word. Hence, this model neither has the representation of MRF nor does it have the ability to perform any kind of inference.

5.3.2 Semantic network nearest neighbor

Another strategy to make predictions is to use the graph structure of the semantic network, but not cast it as a markov random field or perform inference. Instead, for a given target word for which a prediction is to be made, we perform breadth first search using the word

as the source, and find the word in W_{pre} that has the shortest path distance from the target word. Hence, this model has a good representation but lacks the ability to reason using inference.

5.4 Semantics-free models

In this section, I describe models that do not consider word semantics but can be used to make predictions about a child's vocabulary knowledge using some prior ground truth about what the child knows and doesn't know. Namely, I discuss frequency based and phoneme based models.

5.4.1 Frequency based model

As discussed in chapter 2, the theory of incidental learning posits that children often learn new words when they are exposed to them. Further, it suggests that the higher the frequency of exposure to a word, the higher will be the probability of a child committing it to memory [63]. Thus, frequency of word usage in everyday language may be a good indicator of vocabulary knowledge. We use the SUBTLEXus database (74,286 word forms) as a source of word frequency counts of different words used in spoken English language [5]. A common problem with using raw word frequencies to build a model is that frequencies heavily depend on the size of the corpus used. To prevent this problem, we use the zipf scale measure of word frequency counts instead of raw frequency counts. The zipf scale converts word frequencies (per billion words) into a log based scale with values between 1-7 (the higher the frequency, the higher the zipf score) and is independent of the size of the corpus used. Then for each child, we train a personalized logistic regression model using words in W_{pre} as training data. That is, we use the zipf score of each word as a 1-dimensional feature vector and whether or not the child got the word correct as the label (binary classification). Since the dataset used to train the model had a class imbalance (i.e, there were more instances of words known than words unknown), we weighted the classes inversely proportional to their frequency in the training set. The weight w_i for class i is given by:

$$w_i = \frac{n}{kn_i} \tag{5.11}$$

where n is the total number of data points, k is the number of classes (here, $k = 2$) and n_i is the number of instances that have label i. In this way, the model is able to learn the kinds of words that a child knows based on frequency of word occurrence in every language use. While testing, we again find the zipf score of a word and use it as a feature to make predictions about whether or not the child would know a given word.

5.4.2 Phonetics based model

It is known that children retain words when the words are phonetically similar to words they already know. Intuitively, if a child knows the word "cat", it is likely that exposure to words "rat" and "bat" would have aided retention of those words since they are phonetically similar to "cat". A common measure of distance between strings of words is Levenshtein distance which measures the edit-distance of insertion, deletion and replacement of characters between two strings. Here, we use a fixed cost of 1 for each operation. Then, to make a prediction about a given word, we find the nearest word in W_{pre} (using the Levenshtein distance metric) and assign the label of the nearest word. This strategy allows the algorithm to capture the notion of retention of words based on their phonetic similarity using the Levenshtein distance as a proxy measure of phonetic distance.

5.5 Combining semantics-free and semantics-based models

Different cognitive theories posit different ways of how children acquire new words and no one theory alone explains how children learn words. Each model discussed in this chapter is based on a different cognitive theory of learning. In this section, I propose two strategies of combining predictions from different models to evaluate if we can make better predictions by combining the predictive capabilities of each model (R2).

5.5.1 Conditional independence

Let us assume we wish to estimate the probability of a child knowing a given word w, and have two separate models m_1 and m_2 that estimate the probability of the child knowing the word. Thus, we are given $p(w|m_1)$ and $p(w|m_2)$. We then wish to find, $p(w|m_1, m_2)$. If we assume the two models to be conditionally independent sources, then the probability $p(w|m_1, m_2)$ using the bayes optimal method to combine distributions is given by [3]:

$$p(w|m_1, m_2) \propto p(w|m_1)p(w|m_2) \tag{5.12}$$

This method has been used to solve other problems in machine learning as well [21]. In this way, two models that make assumptions using different psycholinguistic theories of learning can be combined to give the final posterior probability of the child knowing a given word.

5.5.2 Mixture of distributions

Another common method of combining probabilities is to create a new distribution that is a mixture of two distributions (weighted sum). Since there is no prior informing which of the two distributions should be given a higher weight, we assume an equal weight (0.5) for each of the distributions. Thus, for each word, the final posterior is given by the weighted sum of distributions from the two models.

5.6 Evaluation, Results and Analysis

In the previous sections, I discussed the mathematical and computational formalism of different models as well as the theories of learning they base their assumptions on. In this section, I first discuss the evaluation methodology used to compare each of the models performance to answer R1 and R2. I then present results and provide an analysis of the results presented. I finally discuss the advantages and limitations of using the models presented in the previous sections.

5.6.1 Evaluation

The research questions R1 and R2 can be seen as binary classification problems. That is, we want to predict the posterior probability of a child knowing some target word given knowledge about the words they already know. To this end, I use each of the models to make predictions about words in *(i)*$W_{post,selected}$ for children in the *(ii)* baseline condition. These two choices are justified as follows:

Selection of $W_{post,selected}$: For any given child, not all words in W_{post} are semantically related to words in W_{pre}. By definition, a model that tries to make predictions about semantically related words cannot make predictions about words that are not semantically related. This occurs in the case where $G_{semantics}$ is a disjoint graph and there exists a word $w_{i,post}$ in an independent subgraph in which there exist no words from W_{pre}.

For example, consider the hypothetical case where a human tutor only knows about a child's knowledge of words related to the solar system (earth, mars, moon, etc). Given this knowledge, they cannot make predictions about words that are related to computers (laptop, keyboard, screen, etc) using only word semantics. Semantics-free models however have the ability to make predictions about any given word irrespective of their semantic relations. These limitations are discussed in detail in subsection 5.6.3.

Thus, in order to compare the models, we must use the set of words for which both kinds of models can make predictions. Moreover, since R1 only asks questions about words in W_{post} that are semantically related to words in W_{pre}, words in $W_{post,selected}$ are used as the test set.

Selection of baseline condition: The baseline condition is the only condition where children did not interact with a robot or go through a learning intervention. In the absence of an intervention, knowledge about W_{pre} can still give information about words in $W_{post,seleted}$. Since, the model is only using words from the pre-test to make predictions, it is unaware of additional knowledge that a child might have gained during the intervention. Thus, if a learning intervention is successful, a child will commit new words to memory and might learn words that they did not know prior to the intervention. In such a case, words

in the pre-test would no longer predictive of words in the post-test.

To further elucidate this reasoning, let us again consider the hypothetical case where a tutor is trying to make predictions about a child's knowledge of the target word "Jupiter". If they observe that the child does not know "Venus", "Saturn" or any of the planets of the solar system, they would probably guess that the child does not know the target word. This set of observations can be seen as analogous to the pre-test. However, after the observations are made, the child now partakes in a learning intervention and learns about most of the planets in the solar system. At this point, the tutor's observations about the child's knowledge prior to the intervention are no longer predictive of the child's knowledge after the intervention. Now, during the post-test, the child might get the word "Jupiter" correct, but the tutor's (or model's) prediction would only be based on knowledge assessed in the pre-test. Thus, we only use the baseline condition to evaluate R1 and R2. In the next chapter, I show how words used in an intervention along with words from the pre-test may be used in making predictions about words in $W_{post,selected}$ for children in the personalized and non-personalized conditions.

A common method of evaluating probabilistic binary classifiers is area under the curve (AUC) of the receiver operating curve (ROC) as it tries to balance true positive and false positive rates by considering a number of different thresholds. However, table 5.1 shows that the dataset in all three conditions is imbalanced. In such cases, area under receiver operating curve can provide a skewed picture of the model's performance. Hence, we compute area under the precision-recall curve which is a more balanced metric for datasets with significant class imbalances [59, 13].

To summarize, we evaluate each model on words from $W_{post,selected}$ for children in the baseline condition by computing AUC for the precision-recall curve per child.

Table 5.1: Number of positive and negative samples per condition

Condition	Number of positive samples	Number of negative samples	Total samples
Personalized	652	277	929
Non-Personalized	536	243	779
Baseline	506	197	703

5.6.2 Results and analysis

Figure 5-3 and table 5.2 show the mean area under the precision-recall curve for each of the models for children in the baseline condition. We find that that among semantics-based models, MRF ($AUC \approx 0.80$) outperforms baseline models. Further, we find that the semantic network representation ($AUC \approx 0.73$) has a higher predictive power when compared to a model that finds the nearest semantically similar neighbor ($AUC \approx 0.71$). More interestingly, the results show that the real power of an MRF comes from its ability to observe nodes and perform inference. The two observations: *(i)* significant increase in AUC from ≈ 0.73 (nearest neighbor in semantic network) to ≈ 0.80 (MRF) and the fact that *(ii)* MRF puts no constraints on path distance between an observed node and a target node, show that the posterior probability of knowing a word is determined by observations of all nodes in W_{pre} in the graph and not just the nearest neighbor. Between the semantics-free models, the frequency based model performed significantly better ($AUC \approx .78$) than the phonetics based model ($AUC \approx 0.73$). It is however surprising that the phonetics based model performed almost as well as the semantics network nearest neighbor model. When comparing the frequency based model with MRF, we find that MRF has a better performance in predicting words in $W_{post,selected}$. It is interesting to note that MRF is able to perform well using inferences based on word similarities alone without any information about how often a child might have been exposed to a given target word.

Subject level analysis: Since MRF is slightly better than the frequency based model as evaluated using mean areas under the precision-recall curve, a subject level analysis allows us to find subjects for which either of the models is better in predicting knowledge of words in $W_{post,selected}$. Table 5.3 shows that out of seventeen subjects, MRF performed

Table 5.2: Mean area under the precision-recall curve and standard deviation (baseline condition) for different models

Model	Mean AUC	Std
MRF	**0.799**	**0.089**
Semantic network NN	0.734	0.071
GloVe nearest neighbor	0.714	0.065
Frequency based	0.782	0.096
Phonetics based	0.730	0.077

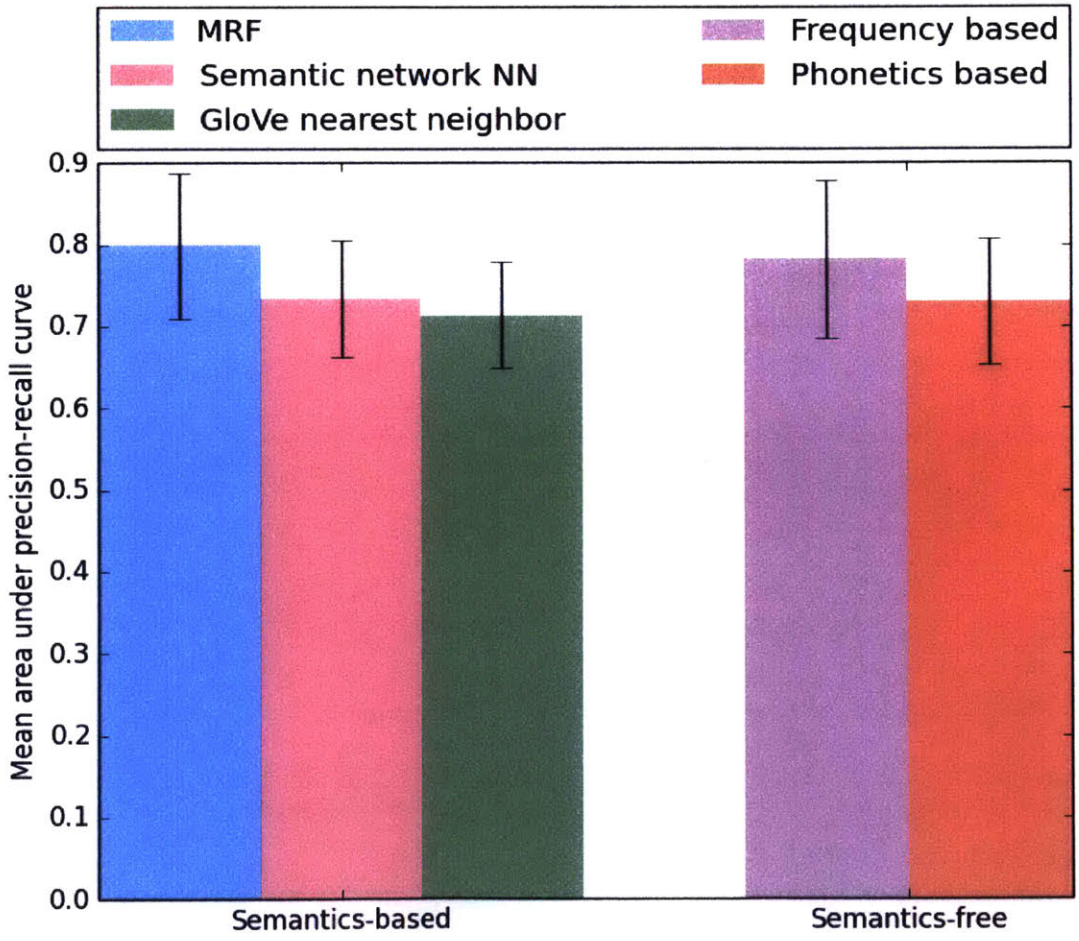

Figure 5-3: Mean area under the precision-recall curve (baseline condition)

Table 5.3: Subject-level analysis of mean area under the PR curve and standard deviation

Condition	Number of subjects	AUC_{mrf}	AUC_{freq}
$AUC_{mrf} > AUC_{freq}$	**11**	$\mathbf{0.819 \pm 0.093}$	0.767 ± 0.11
$AUC_{mrf} < AUC_{freq}$	6	0.762 ± 0.064	$\mathbf{0.81 \pm 0.055}$

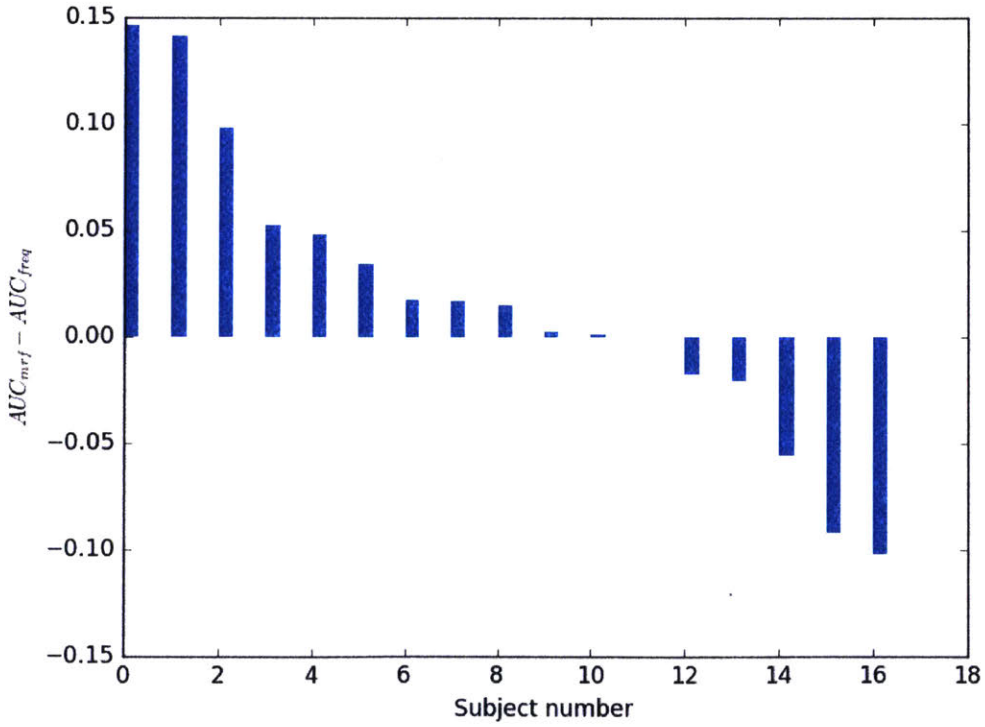

Figure 5-4: Subject level differences between AUC_{mrf} and AUC_{freq}

better than the frequency based model on eleven subjects. Moreover, we observe that the mean AUC scores on the respective set of subjects are similar for each of the conditions given in table 5.3. More concretely, on the set of subjects where MRF dominates, $AUC_{mrf} \approx 0.82$ and $AUC_{freq} \approx 0.76$, while on the set of subjects where the frequency based model performs better, $AUC_{freq} \approx 0.81$ and $AUC_{mrf} \approx 0.76$. Figure 5.3 shows the difference $AUC_{mrf} - AUC_{freq}$ per subject. Here, we further see that there is a distinct set of subjects where each of the models performs significantly better than the other. This finding suggests that while MRF performs better on majority of the subjects, further investigation is required to identify the exact conditions under which the frequency based models should be used instead of MRF. One hypothesis is that, for the set of subjects where the frequency based model performed better, the subjects might have been exposed to new words before taking the post-ppvt. In such a case, words in W_{pre} would no longer be predictive of words in $W_{post,selected}$.

Thus, in reference to the first research question *"R1: Given a child's knowledge about words assessed in the pre-test, can we we predict whether a child would know other semantically related words as assessed in the post-test using a model that uses a theory of mind framework?"*, the aforementioned results show that MRF makes predictions about semantically related words in post-test using words in the pre-test as observations with a reasonable performance. Further, the assumptions made by the model mimic the theory of mind framework.

Combining predictions: Given that both MRF and the frequency based model perform well of different sets of subjects, I evaluate if combining predictions enables increases the performance of either of the models. Table 5.4 shows that there is a slight improvement in prediction performance for MRF ($\approx 0.3\%$) and a greater improvement in performance for the frequency based model ($\approx 2\%$). Both strategies of combining predictions give similar gains in improvement. This improvement is seen even though the set of children where the two models perform well are mutually exclusive and the difference in performance is significant as seen in figure 5-4. These results show that combining two models that are based on different theories of learning can help improve prediction performance. It is important to

Table 5.4: AUC after combining predictions

Model	Mean AUC	Std
MRF	0.799	0.089
Frequency based	0.782	0.096
Combined (conditional independence)	0.802	0.09
Combined (mixture of distributions)	**0.803**	**0.09**

note that the predictions were combined using uniform priors for both methods. While the strategies to combine predictions work, the improvement in not significant (in the case of mrf) because of the absence of informative priors. Therefore, given that each model performs well on a different set of subjects and the fact that predictions can be combined using the aforementioned strategies, pre-determining the weights (*priors*) for each model per subject would further help make better predictions about a child's knowledge.

Thus, in reference to the second research question *"R2: Can we use semantics-free models in conjunction with the semantics-based model to make better predictions about words assessed in the post-test?"*, results show that it is possible to make better predictions by combining the predictions of individual models (MRF and frequency based) with a caveat that more informative priors would significantly help in combining predictions to achieve better performance.

5.6.3 Advantages and limitations of semantics-based models

Since the prediction performance of MRF was the highest among semantics-based models and the performance of the frequency-based model was highest among semantics-free models, in this section, I compare the advantages and limitations of those two models. One of the main limitations of a semantics-based model (MRF) is that it can only make predictions about words that are *semantically related* (as defined in 5.1.3) to words for which it has some information. On the other hand, since the frequency based model only takes frequency of word exposure into account, it can make predictions about any word (since every word has some frequency of occurring in everyday language). Despite these limitations, we have seen that on words that are semantically related to *known words*, MRF performs better than

the frequency based model. However, the main advantage of using a semantics-based model over a frequency-based model can be seen by analyzing the assumptions and workings of the models individually.

Each of these models tries to predict the probability of a child knowing a word. In other words, they try to predict the probability that a child might have learned a given word in the past. MRF tries to identify how a target word *fits* with respect to other words that a child already knows in the implicit semantic network representation of a child's knowledge. Thus, it makes predictions about how a child might have learned a new word in the context of other words. A frequency-based model fundamentally makes an assumption that the child has been exposed to a given word and uses the frequency of the word's usage in everyday language as a proxy for how often the child might have been exposed to the given word. However, when *teaching* a child a new target word (a word that a child has not been exposed to before) a frequency based model cannot make any predictions about how well the child is prepared to learn that word. A semantics-based model can however be used to probe the child about semantically related nodes and make predictions about how well the word *fits* in the child's semantic network, thus predicting the probability of how well the child might learn the word. Moreover, as a bi-product of performing LBP on the graph, MRF also computes the probability of knowing other semantically related words in the graph that the child was not assessed for. As discussed in chapter 2, teaching a child words that are semantically similar to words they already know improves retention and recall of new words [69, 14]. These advantages make a semantics-based model suitable not only for predicting existing knowledge but also for developing novel intervention strategies.

Chapter 6

Applications of semantics-based model

In the previous chapter, we showed how MRF can be used to make predictions about a child's existing existing vocabulary using words in W_{pre} for children in the baseline(B) condition. As discussed in chapter 4, in the Personalized (P) and the non-personalized (NP) conditions, the robot narrated a story to children with the purpose of teaching new words to them. After listening to each story, children were asked to retell the story back to the robot. In the personalized condition, the robot told stories that maximized learning outcomes while keeping children engaged in the story. In the non-personalized condition, the robot did not personalize content to increase children's learning outcomes.

In this chapter, I show how we can make predictions about a child's vocabulary by using words that the robot used in the stories it narrated ($W_{robot,story}$). These words give a direct observation of words that each child heard during the intervention. Additionally, I show how we can use words from a child's retell of the story ($W_{child,story}$) to make predictions about their vocabulary knowledge. These words provide a direct observation of the words that a child might have learned as a result of listening to new words used in a story. Finally, I show how these results can be used to determine how successful an intervention was and answer the third research question *"R3: Given data about stories that the robot narrated and data from a child's story retell, can we use the sematics-based model to deter-*

mine how well a child learned new words in the context of other semantically related words?".

I first describe the method used to add new words to the semantic network $G_{semantics}$ for each child and provide details about the experiment used to answer R3. I then describe the evaluation methodology, present the results and provide an analysis of the results in the context of R3.

6.1 Method and evaluation

In order to add new words from the storytelling intervention, we augment graph $G_{semantics}$ with words used in the intervention. When we add words to the network from the robots narration of stories, we use words from $W_{robot,story}$ and when we add words from a child's story retell, we use words from $W_{child,story}$. For the purpose of generalization, let the set of words $W_{intervention}$ represent either of the sets of words depending on which words we would like to augment the graph with. For a given word w in $W_{intervention}$, we find all the words in $G_{semantics}$ that are semantically similar to w. Let us call this set of words $W_{similar}$. We then add an edge between w and all the words in $W_{similar} \subseteq V_{semantics}$. In this way, we augment the graph $G_{semantics}$ with new words used in the intervention. Since, MRF is constructed from $G_{semantics}$, new nodes are similarly added to G_{mrf}. We then perform inference on G_{mrf} in the same way as described in chapter 5.

In the previous chapter, we defined potentials for a pair of words as an exponential function of energy. The potential function was associative in its form and the energy of a pair of words was further parameterized by the semantic similarity between words. The equation for the potential function captured the assumption that if a child knows a word, then they would know other nodes that are semantically similar to the given word. However, for the new words added from $W_{intervention}$, we wish to evaluate if the assumption is true or not. That is, we wish to identify if a child indeed added the new nodes to their semantic network. Hence, for words in $W_{intervention}$, the pairwise potentials represent how well a child associated a new word with existing words in $G_{semantics}$. Thus, for words in $W_{intervention}$,

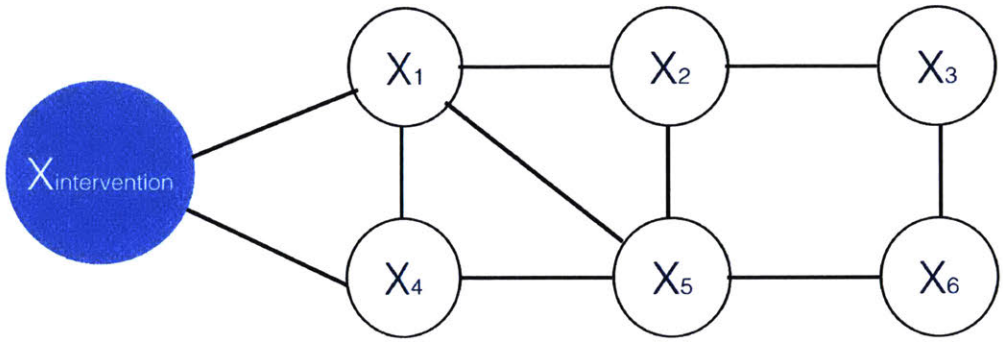

Figure 6-1: Augmenting graph from words used in intervention

energy of a pair of words is parameterized by a function that represents how well the child learned a given word. Let us call this function $l(w_1, w_2)$. We keep the same associative and exponential form of the potential function and define it as follows:

$$\phi(X_i, X_j) = \begin{bmatrix} e^{-(1-l(w_i,w_j))} & e^{-l(w_i,w_j)} \\ e^{-l(w_i,w_j)} & e^{-(1-l(w_i,w_j))} \end{bmatrix} \tag{6.1}$$

Further, we restrict the value of $l(w_1, w_2) \geq 0.5$. For values < 0.5, the potential function no longer enables the MRF to remain associative. Intuitively, a value < 0.5 would mean that if a child learned a given word, then they would not know the words connected to it by an edge. A value of 0.5 means that observing the node would have no effect on the conditional marginals of other nodes, indicating that the word was not learned.

If a child successfully learned words in $W_{intervention}$ while associating them with other semantically similar words, then making positive observation on these nodes should result in a better prediction performance on words in $W_{post,selected}$, since the model makes correct observations about the child's true knowledge. However, if we assume that a child learned new words by using the words in $W_{intervention}$ as positive observations, but the child did not

actually commit the words to their memory, then we would see a decrease in the model's performance on words in $W_{post,selected}$, since the model makes incorrect observations about the child's true knowledge.

Thus, in order to evaluate if a child learned new words from the intervention, we first perform inference using words only in W_{pre}. We then augment the graph $G_{semantics}$ with words in $W_{intervention}$ and make an assumption that children learned these words as a result of the intervention. We then perform a parameter search by assuming different values of $l(w_1, w_2)$ and perform inference. If we see an increase in performance of the model for any $l(w_1, w_2) > 0.5$, we say that our assumption is correct and that the child successfully associated the given words with the words in $G_{semantics}$. However, if we observe a decrease in performance, or find that $l(w_1, w_2) = 0.5$ gives the best performance, then our assumption is incorrect and we say that the child did not successfully commit most of the words to memory. We again use mean AUC as a measure of the models performance. Since, $l(w_1, w_2)$ is a parameter of the energy function, I refer to this parameter as k in the next section for ease of terminology. We performed this experiment for children in personalized and non-personalized conditions using words from $W_{robot,story}$ and $W_{child,story}$ for each condition.

Another strategy to check if children learned words from the intervention is to check children's performance on W_{common} (explained in Chapter 4), since this set of words was used during the intervention for both personalized and non-personalized conditions and was assessed both prior to and after the intervention for each child. In this way, W_{common} may be seen as a proxy for all the words used in the intervention. The assumption we make here is that, *if children were able to learn this set of words as a result of the intervention, then they must have learned other words also used in the intervention.*

6.2 Results and analysis

In this section I present and discuss the results of experiments described in the previous section. I first provide an analysis of experiments run on data from children in the personalized

Table 6.1: Mean area under the precision-recall curve and standard deviation for observations of different sets of words for children in the personalized condition (optimal $k = 0.65$)

Condition	Mean AUC	Std
Pre-test	0.75	0.129
Child' story retell	0.753	0.119
Child's story retell + Pre-test	0.768	0.129
Robot's story	0.767	0.122
Robot's story + Pre-test	**0.776**	**0.120**

condition. I then follow it up with a similar analysis on data collected from children in the non-personalized condition. Finally, I corroborate the conclusions drawn from the preceding experiments with a separate analysis of words in W_{common}.

6.2.1 Personalized condition

Table 6.1 shows that when only words from the pre-test (W_{pre}) were used to make predictions of words in $W_{post,selected}$, the model's mean $AUC \approx 0.75$. Augmenting graph $G_{semantics}$ with words used in the robot's story $W_{robot,story}$ and making observations for words used in W_{pre} in addition to making positive observations for words in $W_{robot,story}$ yields an increase in the model's performance ($AUC \approx 0.78$). The optimal value of the parameter was computed to be $k = 0.65$ for all conditions reported in 6.1. This result indicates that our assumption that children learned majority of the words used in the *robot's stories* is correct. We then followed the same process, but used words from children's story retell ($W_{child,story}$) instead of words in $W_{robot,story}$. In this case, the model still performed better ($AUC \approx 0.77$) than the condition where only words from W_{pre} were used, indicating that children indeed understood the words they used to *retell* the robot's stories. However, the model's performance was slightly lower in this condition when compared to the condition where words from $W_{robot,story}$ were used. This result is evidence of the fact that a children might have learned words from the robot's stories that they did not explicitly use while retelling the stories. Finally, we compare each of the two conditions using only words in $W_{robot,story}$ and $W_{child,story}$ without making observations for words in W_{pre} and find a decrease in the model's performance. This decrease in performance shows that words in W_{pre} provide *useful* ground truth information

about a child's vocabulary before the intervention and hence benefit the model in making better predictions about a child's vocabulary.

6.2.2 Non-personalized condition

Table 6.2: Mean area under the precision-recall curve and standard deviation for observation of different sets of words for children in the non-personalized condition (optimal $k = 0.5$). $k = 0.65$ is used to report results for all sets of words except pre-test.

Condition	Mean AUC	Std
Pre-test	**0.772**	**0.127**
Child' story retell	0.738	0.130
Child's story retell + Pre-test	0.761	0.130
Robot's story	0.728	0.120
Robot's story + Pre-test	0.742	0.119

For the analysis of non-personalized condition, I used the same evaluation methodology as the personalized condition. Table 6.2 shows that when only words from the pre-test (W_{pre}) were used to make predictions about words in $W_{post,selected}$, the model's mean $AUC \approx 0.77$. Further, the optimal value of k was found to be 0.5 for all sets of words indicating that for any $k > 0.5$, the model's performance decreased. This result indicates that the assumption that children learned majority of the words used in the robot's story or in their retelling of the story, is incorrect. Thus, even though children used words from the robot's stories, they did not associate the words with other semantically similar words. Upon empirical investigation of the stories, we further found that children in the non-personalized condition told shorter and less coherent stories as compared to children in the personalized condition. Even though the optimal parameter was found to be $k = 0.5$, I report values for $k = 0.65$ for all other sets of words to show other trends that exist in the data. These trends were consistent across different values of $k > 0.5$. Contrary to the trend in the personalized condition, using child's *story retell + pre-test* yielded higher mean AUC when compared to *robot's story + pre-test.* One hypothesis to explain this result is that while retelling stories, children used words that were indicative of their own vocabulary knowledge rather than using words that they heard in the robot's story. Further, we find that making observations for words in $W_{intervention}$ along with words in W_{pre} gives better performance when compared

to only using words in $W_{intervention}$. This observation is consistent across both personalized and non-personalized conditions indicating that ground truth knowledge of words in W_{pre} aids the model in making better predictions about a child's vocabulary.

The above results show that the personalized condition was better suited for children to learn new words (in the context of other semantically related words) than the non-personalized condition.

6.2.3 Analysis using common words

Since the set of words W_{common} is assessed in all conditions before and after the intervention, we may use children's assessment on these words as a proxy to measure the increase in children's vocabulary in each condition. Figure 6-2 shows children's performance on W_{common} before and after the intervention in each condition. Children in all three conditions had similar performance on the pre-test and show some increase in vocabulary in the post-test. Children in the personalized condition showed maximum increase in vocabulary followed by children in the non-personalized condition. Moreover, the difference between the median vocabulary levels of the two conditions was found to be statistically significant ($p < 0.05$). These results further corroborate the results presented in the previous sections showing that the personalized condition was better suited to enhance children's learning outcomes.

Thus, in reference to the third research question *"R3: Given data about stories that the robot narrated and data from a child's story retell, can we use the sematics-based model to determine how well a child learned new words in the context of other semantically related words?"*, the aforementioned results show that words from both sources of information, ground truth from pre-test and words in $W_{intervention}$ together can be used by an MRF to *(i)* make predictions about children's vocabulary and *(ii)* evaluate whether children learned new words (through a storytelliing intervention) in the context of other semantically similar words.

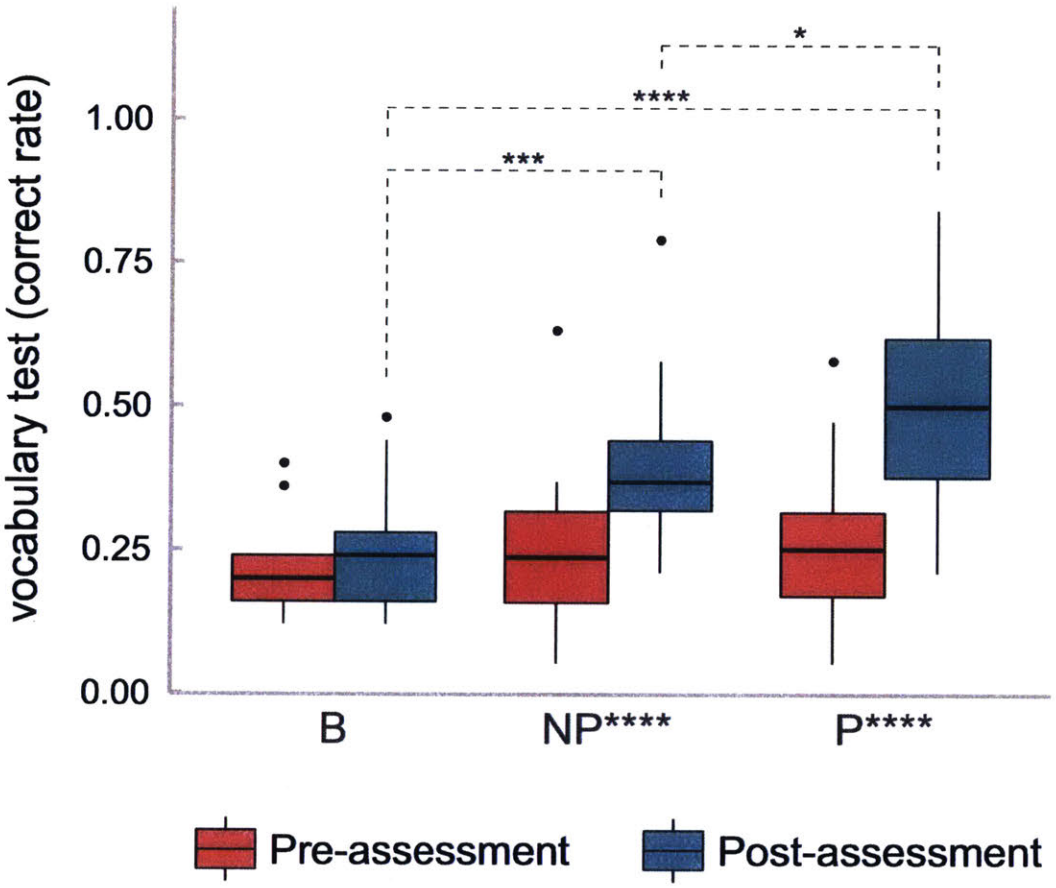

Figure 6-2: Pre-assessment and Post-assessment results on words in W_{common}

Chapter 7

Conclusion

7.1 Contributions

As humans, we make inferences about other people's knowledge from partial information with ease. We especially perform this kind of inference when we assume the role of tutors. Several studies in psycholinguistics have shown that children learn new words by forming semantic relationships with words they already know. In fact, learning the meaning of a new word entails learning how it is semantically related with other words and how it may be used in everyday language. Before providing new information to a tutee, human tutors often implicitly use semantics to draw inferences about the tutee's knowledge.

In this thesis, I use the domain of word learning and present a cognitively inspired semantics-based model to make inferences about a child's vocabulary from partial information about their existing vocabularies, akin to how a human tutor would. I presented results using semantics-free models to make predictions and show that while the semantics-based model performs best *on the whole*, different models perform well for different children. I further discussed the advantages and disadvantages of using each of the presented models for making predictions. Second, I presented two methods to combine predictions from semantics-based and semantics-free models and show that combining predictions can improve predictions about a child's vocabulary knowledge. Finally, I presented an application of the semantics-based model to evaluate if an intervention was successful in teaching children new words. More concretely, I showed that a personalized word learning intervention

with a robotic tutor (that tries to maximize engagement and learning) is better suited for enhancing children's vocabulary when compared to a non-personalized intervention. These results motivate the use of semantics-based models not only in predicting a child's vocabulary knowledge but also in building interventions for teaching children new words such that they are able to understand the semantic meaning behind words.

7.2 Future work

A tutor performs two explicit tasks when providing information: *(i)* make predictions about the tutee's existing knowledge and *(ii)* provide new knowledge in the best possible manner for the tutee to learn. The proposed model performs the first task by analyzing how a given word is *situated* in a child's semantic knowledge representation (semantic network). It then makes predictions about whether or not a child would know a given word based on this information. To perform the second task, the same model can be extended to identify the words that would be easiest for a tutee to learn based on partial knowledge of what they already know. Thus, the semantics-based model may be extended to find *trajectories* or *sequence* of words that should be taught in order to teach a given corpus of words.

Future work will investigate how the model can be used by a robotic tutor to incorporate an active-learning feedback loop where a tutor teaches new words based on inferences and gets feedback on whether the child learned those words. The tutor would then update its parameters from the new "ground truth" data to improve its estimate and again teach new words based on the updated parameters. Further research in encoding better *priors* about a child's existing semantic network would also enhance the models performance. The model presented in this thesis also has the potential to find applications in domains that extend beyond predicting a child's vocabulary knowledge. For example, conversational agents could use the model to find other semantically related concepts that a human might know or find interesting by forming a similar semantic network of concepts and performing inference using past conversations as observations.

Bibliography

[1] Rhyme and alliteration, phoneme detection, and learning to read.

[2] John R Anderson and Michael Matessa. Explorations of an incremental, bayesian algorithm for categorization. *Machine Learning*, 9(4):275–308, 1992.

[3] C Bailer-Jones and K Smith. Combining probabilities. *Data Processing and Analysis Consortium (DPAS), GAIA-C8-TN-MPIA-CBJ-053*, 2011.

[4] Victor Bandeira de Mello, G Bohrnstedt, C Blankenship, and D Sherman. Mapping state proficiency standards onto naep scales: Results from the 2013 naep reading and mathematics assessments. nces 2015-046. *National Center for Education Statistics*, 2015.

[5] Marc Brysbaert and Boris New. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990, 2009.

[6] Susan Carey. *The child as word learner*. na, 1978.

[7] Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, pages i–174, 1998.

[8] Noam Chomsky. *Language and Mind*. Cambridge University Press, 2006.

[9] Molly F Collins. Ell preschoolers? english vocabulary acquisition from storybook reading. *Early Childhood Research Quarterly*, 25(1):84–97, 2010.

[10] Eliana Colunga and Linda B Smith. From the lexicon to expectations about kinds: A role for associative learning. *Psychological review*, 112(2):347–382, 2005.

[11] Robert Conrad and Audrey J Hull. Information, acoustic confusion and memory span. *British journal of psychology*, 55(4):429–432, 1964.

[12] David E Copeland and Gabriel A Radvansky. Phonological similarity in working memory. *Memory & Cognition*, 29(5):774–776, 2001.

[13] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[14] Yeşim Dilek and Nurcihan Yürük. Using semantic mapping technique in vocabulary teaching at pre-intermediate level. *Procedia-Social and Behavioral Sciences*, 70:1531–1544, 2013.

[15] Lloyd M Dunn and Douglas M Dunn. *PPVT-4: Peabody picture vocabulary test.* Pearson Assessments, 2007.

[16] John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

[17] Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, 20(5):578–585, 2009.

[18] Edward Fry. The new instant word list. *The Reading Teacher*, 34(3):284–289, 1980.

[19] Goren Gordon and Cynthia Breazeal. Bayesian active learning-based robot tutor for children's word-reading skills. In *AAAI*, pages 1343–1349, 2015.

[20] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3951–3957. AAAI Press, 2016.

[21] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*, pages 2625–2633, 2013.

[22] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.

[23] Thomas L Griffiths, Edward Vul, and Adam N Sanborn. Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21(4):263–268, 2012.

[24] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[25] Betty Hart and Todd R Risley. *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, 1995.

[26] Susan J Hespos. Language acquisition: when does the learning begin? *Current Biology*, 17(16):R628–R630, 2007.

[27] Jan H Hulstijn et al. Incidental and intentional learning. *The handbook of second language acquisition*, pages 349–381, 2003.

[28] Larry L Jacoby and Mark Dallas. On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, 110(3):306, 1981.

[29] Dale D Johnson, Susan Toms-Bronowski, and Susan D Pittelman. An investigation of the effectiveness of semantic mapping and semantic feature analysis with intermediate grade level children (program report 83-3). madison: University of wisconsin. *Wisconsin Center for Educational Research*, 1982.

[30] Susan S Jones, Linda B Smith, and Barbara Landau. Object properties and knowledge in early lexical learning. *Child development*, 62(3):499–516, 1991.

[31] Michael I Jordan and Yair Weiss. Probabilistic inference in graphical models. *Handbook of neural networks and brain theory*, 2002.

[32] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[33] Roger P Levy, Florencia Reali, and Thômas L Griffiths. Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in neural information processing systems*, pages 937–944, 2009.

[34] Jacques Mehler, Peter Jusczyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, and Claudine Amiel-Tison. A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178, 1988.

[35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[36] SungJin Nam, Gwen Frishkoff, and Kevyn Collins-Thompson. Predicting short-and long-term vocabulary learning via semantic features of partial word knowledge. *Ann Arbor*, 1001:48109, 2017.

[37] Kate Nation and Margaret J Snowling. Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of memory and language*, 39(1):85–101, 1998.

[38] Thierry Nazzi, Josiane Bertoncini, and Jacques Mehler. Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, 24(3):756, 1998.

[39] Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson. A cognitive model of semantic network learning. In *EMNLP*, pages 244–254, 2014.

[40] Mariela M Páez, Patton O Tabors, and Lisa M López. Dual language and literacy development of spanish-speaking preschool children. *Journal of applied developmental psychology*, 28(2):85–102, 2007.

[41] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. Telling stories to robots: The effect of backchanneling on a child's storytelling. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 100–108. ACM, 2017.

[42] Hae Won Park, Rinat Rosenberg-Kima, Maor Rosenberg, Goren Gordon, and Cynthia Breazeal. Growing growth mindset with a social robot peer. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 137–145, New York, NY, USA, 2017. ACM.

[43] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[44] Steven Pinker and Alan Prince. Regular and irregular morphology and the psychological status of rules of grammar. *The reality of linguistic rules*, 321:51, 1994.

[45] Melissa Allen Preissler. Associative learning of pictures and words by low-functioning children with autism. *Autism*, 12(3):231–248, 2008.

[46] David Premack. Is language the key to human intelligence? *Science*, 303(5656):318–320, 2004.

[47] Willard Van Orman Quine. *Word and object.* MIT press, 2013.

[48] Tina Rosenberg. The power of talking to your baby. *Power*, 3:25, 2013.

[49] DC Rubin. Schemata: The building blocks of cognition. *Theoretical issues in reading comprehension, ed. RJ Spiro, BC Bruce, WF Brewer. Erlbaum.[aAK]*, 1980.

[50] Jenny R Saffran, Ann Senghas, and John C Trueswell. The acquisition of language by children. *Proceedings of the National Academy of Sciences*, 98(23):12874–12875, 2001.

[51] Adam N Sanborn, Thomas L Griffiths, and Daniel J Navarro. Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, 117(4):1144, 2010.

[52] Edward Sapir. *Culture, language and personality: Selected essays*, volume 342. Univ of California Press, 1985.

[53] Hollis S Scarborough. Index of productive syntax. *Applied psycholinguistics*, 11(1):1–22, 1990.

[54] Alan Searleman and Douglas J Herrmann. *Memory from a broader perspective.* McGraw-Hill New York, 1994.

[55] Jeffrey Mark Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91, 1996.

[56] Burrhus Frederic Skinner. *Verbal behavior.* BF Skinner Foundation, 2014.

[57] Kenny Smith, Andrew DM Smith, and Richard A Blythe. Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35(3):480–498, 2011.

[58] Linda B Smith. How to learn words: An associative crane. *Breaking the word learning barrier*, 2000.

[59] Samuel Spaulding, Huili Chen, Safinah Ali, Michael Kulinski, and Cynthia Breazeal. A social robot system for modeling children's word pronunciation: Socially interactive agents track. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1658–1666. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[60] Samuel Spaulding, Goren Gordon, and Cynthia Breazeal. Affect-aware student models for robot tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 864–872. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[61] Sebastian P Suggate, Wolfgang Lenhard, Elisabeth Neudecker, and Wolfgang Schneider. Incidental vocabulary acquisition from stories: Second and fourth graders learn more from listening than reading. *First Language*, 33(6):551–571, 2013.

[62] Joshua B Tenenbaum. Rules and similarity in concept learning. In *Advances in neural information processing systems*, pages 59–65, 2000.

[63] Feng Teng. The effects of word exposure frequency on incidental learning of the depth of vocabulary knowledge. *GEMA Online® Journal of Language Studies*, 16(3), 2016.

[64] Paul Thagard. *Mind: Introduction to cognitive science*, volume 4. MIT press Cambridge, MA, 1996.

[65] John C Trueswell, Irina Sekerina, Nicole M Hill, and Marian L Logrip. The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, 73(2):89–134, 1999.

[66] Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

[67] M Westerveld and GT Gillon. Language sampling protocol. *Christchurch: University of Canterbury*, 2002.

[68] Benjamin Lee Whorf. Language, thought and reality. selected writing: of benjamín lee whorf. *Cambridge, MA.: The MIT Press*, 1956.

[69] Maryanne Wolf, Mirit Barzillai, Stephanie Gottwald, Lynne Miller, Kathleen Spencer, Elizabeth Norton, Maureen Lovett, and Robin Morris. The rave-o intervention: Connecting neuroscience to the classroom. *Mind, Brain, and Education*, 3(2):84–93, 2009.

[70] Chen Yu and Dana H Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165, 2007.