A Computational Model of the Acquisition of Mental State Verbs

by

Libby Barak

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

© Copyright 2016 by Libby Barak
Abstract

A Computational Model of the Acquisition of Mental State Verbs

Libby Barak
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2016

Children first use verbs to refer to inner mental states, such as thoughts and beliefs, only towards their 4th birthday. Presenting this ability serves as important evidence of the stage of cognitive development of a child since it entails an ability to conceptualize mental states. This developmental milestone is necessary for children to engage in social interaction since appropriate interaction relies on directing the interaction and anticipate actions surrounding them given the mental states of others. Thus far, psycholinguistic models have been widely focused on the development of cognitive and linguistic skills required to enable the expression of mental states by children. However, while computational models have proved to be a useful tool in studying the developmental trajectory of similar language acquisition phenomena, no computational model to our knowledge has been used to analyze the linguistic development required for the expression of mental states. In this thesis, I present a computational model used to analyze the way linguistic development plays a role in learning to express mental state verbs.

The computational model in this thesis offers an integrated framework that simultaneously models several of the cognitive and linguistic factors in the acquisition of mental state verbs. The cognitive factor simulates the difficulty in attending to mental states, while the linguistic properties represent the typical semantic and syntactic properties of use of mental state verbs, e.g., reference to mental states using a sentential complement syntactic structure. The experimental work in this thesis replicates psycholinguistic
observations from the child acquisition of mental state verbs. The results of these experiments shed light on the facilitating role of certain linguistic properties of mental state verbs in learning to verbally refer to mental meanings. Importantly, I achieve these results within the context of naturalistic language input that mimics the complexity and diversity of use of mental state verbs and verbs from additional semantic classes in child-directed speech. Finally, I present a novel extension to an existing computational model of verb argument structure learning that enables the simultaneous and incremental learning of verb classes. The novel model offers a probabilistic framework to analyze monotonically growing verb classes in comparison to previously offered batch models that limit the capabilities of the computational simulation. Moreover, this model gives support to the importance of an additional linguistic property of mental state verb, i.e., the role of their use with syntactic structures other than the well-studied co-occurrence with sentential complement syntax in child-directed speech. Finally, I show the contribution of this novel computational component in comparison with previous computational models in a wider context of argument structure learning.
Acknowledgements

There are many points during the PhD journey that you feel grateful beyond what words can express. This section gives a unique opportunity to thank the many kind people you met along the way. I can only hope to find the words that would truly express my gratitude.

I was lucky to have Suzanne Stevenson as my supervisor. Suzanne has opened the door to cognitive modeling for me, and cared enough to gently push me forward when it seems I would never get through the opening. She taught me a lot about how to be a good critic of my work, how to tell the story well, and how to always be respectful of the work of others and sincere about the limitations of your own work. But, she also taught me a lot about facing your fears and making the hard decisions. Her guidance helped me become a better researcher and a better mom, for which I am grateful (and unknowingly, so are my kids).

My co-unofficial-supervisor, Afsaneh Fazely has been a mentor, a patient friend, and a true supporter. She taught me a lot about the art of writing, about believing in yourself, and about being a realistic optimist. I’ll always treasure our long discussions about academia, women’s status, and immigration.

I want to thank my committee members, Rich Zemel and Craig Chambers. I have some unforgettable memories of the most awesome questions any student can get in a checkpoint. They inspired me with their ability to provide objective feedback that brings a variety of research domains into play as interdisciplinary research should be. I am thankful to my external examiner, Dan Roth, for his valuable feedback on this dissertation, and for giving me insightful suggestions for future research. I would also like to thank Frank Rudzicz, my internal examiner, for his time and his contagious passion for research that shined through his fascinating questions.

I’m also in debt to many more scholars who had a major role in my academic decisions and achievements. I would like to thank them in chronological order: Ido Dagan, my
M.Sc supervisor, who introduced me to NLP and remained a thoughtful supporter and mentor long after my graduation; Graeme Hirst, who met me in Israel a long time ago and gave me the wonderful opportunity of becoming a UofT student; Sven Dickinson was the first faculty member to give me advice as a UofT professor ("take your time finding a thesis topic since you’ll spend years studying it"), little did he know I would listen to him so carefully; Gerald Penn has not only inspired me with his knowledgeable questions throughout the years, but also made me feel at home by discussing Israelis and plagiarism in a Speech processing course; Paul Gries has inspired me with his excellence in teaching; and Afra Alishahi, who gave me the code to the original model and showed me true kindness and generosity. I’m also grateful to many wonderful researchers who gave me some wonderful feedback on my work and inspired me with their brilliant comments and suggestions during conferences and talks: Vera Demberg, Adele Goldberg, Daphna Heller, Shalom Lappin, Roger Levy, David Reitter, and Thamar Solorio.

I was also fortunate to be surrounded by some brilliant students. I would like to thank all members of the CL group over the years for the supportive and friendly research environment. I would like to mention, Chris Parisien who showed me true Canadian kindness, I owe this PhD partially to him; Aida Nematzadeh was the first friendly face I’ve met at UofT and has been a true friend ever since; Jackie Chi Kit Cheung has inspired me with his effortless wisdom; Tong Wang has taught me a lot about overcoming challenges in academic research and in facing cultural differences; Varada Kolhatkar has shown me the power of inner peace and positive energy; Katie Fraser inspired me with her ability to balance life and career while excelling in both; Nona Naderi has brightened my days with her genuine smile; Barend Beekhuizen brought Europe to our office; Amin Tootoonchian, although not officially a CL student, has been the best conference buddy ever. I’m also grateful to my CL colleagues from the M.Sc studies, Shachar Mirkin, Eyal Shnarch, and Idan Szpektor, for their continuous support, care, and valuable advice.

Knowing I can’t thank each and every person I’ve met over the years in Toronto, I
would like to thank some of the people who made this journey much more pleasant. I would like to thank Pamela Lenkov, Relu Patrascu, Luna Boodram, Lisa DeCaro, Marina Haloulos, and Vinita Krishnan the wonderful staff at UofT. Fernando Flores-Mangas for his sense of humor and cold beers, every cynical word made me feel right at home even during the worst snow storms; Inamr Givoni for being my official mentor as I came to UofT; Julia and Ophir Vinik, Meital Pinto, Boaz Miller, Meital Avivi-Reich, Dori Reich, Hila Fogel-Yaari, Baruch Fogel, Micha Livne, Alex Levinshtein, Sharon Pauker, Ella and Carmi Daniel, and Shira Agasi have been my Israeli community away from home. I thank them for every holiday dinner and Hebrew word of encouragement. I also thank the friends who supported from a distance, Oshrit, Avital, Rotem, Tzvia, Michael, Boaz, Avital and Haim. As many international students pointed out to me over the years, Israelis have this unique ability to become instant friends if only for the common ethnicity. I would like to thank Yoav Goldberg, Dan Goldwasser, and Yuval Marton, among others for the joyful moments we shared during some ACL conferences.

Finally, but most importantly, I would like to thank my family. I’m of the few lucky ones to have the most amazing parents. Only now, being a parent myself, I understand the sacrifices and efforts put into raising the four of us. I appreciate every word of encouragement, every thoughtful piece of advice, every visit to Toronto, every hug and kiss, and every brother. Oshik, Ran, and Ido, as mom always said, friends come and go, but you always had my back. I thank my in-laws, Edna, Israel, Meital, and Matan Barak for their patience and support over this long journey so far away from home. Oren, we set off on this journey not knowing what lies ahead of us. I’m grateful for all the moments we’ve shared and look forward to what’s ahead. Lastly, the wind beneath my wings, my inspiration and joy, Guy and Yoav - you taught me more about language acquisition than any paper could, you taught me about true love and true happiness. Anything I can add to this wouldn’t capture the extent to which I’m grateful for being blessed with each and every member of my family. Toda.
Contents

1 Introduction 1

2 Acquiring Mental State Verbs 8
   2.1 Overview ................................................. 8
   2.2 Challenges in the Acquisition of Mental State Verbs .......... 9
   2.3 Psycholinguistic Studies of MSV Acquisition ................. 14
      2.3.1 The Performative Stage ............................ 15
      2.3.2 The Mental Production Stage ......................... 20
   2.4 Computational Models ................................... 24
      2.4.1 Computational Models of Theory of Mind .............. 26
      2.4.2 Computational Models of verb learning ................ 28

3 Attending to Mental State Verbs 36
   3.1 Introduction .............................................. 36
      3.1.1 Background and Our Approach ........................ 39
   3.2 The Computational Model .................................. 41
      3.2.1 Overview of the Model ................................ 41
      3.2.2 Learning Constructions of Verb Usages ............... 43
      3.2.3 Modeling Developing Attention to Mental Content ..... 46
   3.3 Experimental Setup ....................................... 49
      3.3.1 Input Data ............................................ 49
5.3.2 Experimental Results ........................................ 120
5.4 Discussion ...................................................... 124

6 Conclusions .................................................. 126

6.1 Summary of Contributions ................................... 127
6.2 Future Extensions ........................................... 131

Bibliography .................................................. 137
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>A scene–utterance pair and the frame extracted from it.</td>
<td>42</td>
</tr>
<tr>
<td>3.2</td>
<td>A scene–utterance pair and the two frames extracted from it.</td>
<td>47</td>
</tr>
<tr>
<td>3.3</td>
<td>A sample of the lexical entries for the verb <em>know</em> in our input generation lexicon.</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>The three test scenarios.</td>
<td>52</td>
</tr>
<tr>
<td>3.5</td>
<td>The frame representing a physical interpretation of the scene–utterance pair.</td>
<td>58</td>
</tr>
<tr>
<td>4.1</td>
<td>An example input frame.</td>
<td>66</td>
</tr>
<tr>
<td>4.2</td>
<td>The 13 FCT verbs in our data.</td>
<td>70</td>
</tr>
<tr>
<td>4.3</td>
<td>The 13 FCT verbs in our data given the balanced and the skewed distributions.</td>
<td>80</td>
</tr>
<tr>
<td>4.4</td>
<td>The list of our 31 verbs in the input generation lexicon.</td>
<td>85</td>
</tr>
<tr>
<td>5.1</td>
<td>Frequency data of non-alternating (non-ALT) and alternating (ALT) verbs in our lexicon.</td>
<td>107</td>
</tr>
</tbody>
</table>
List of Figures

3.1 An illustration of the clusters learned by the model. 43
3.2 Likelihood of Belief verb prediction given Action or Belief input. 56
3.3 Verb class likelihood. 58

4.1 Verb repetition accuracy for each of the 13 FCT verbs. 73
4.2 Verb substitution rate for each of the 13 FCT verbs. 74
4.3 An illustration of the skewed and balanced conditions. 78
4.4 The 3 highest likelihood values of semantic properties of the event. 82
4.5 An illustration of the frequency of each of the 5 verb classes with SC syntax. 86
4.6 Verb prediction results by human and computational simulations. 88
4.7 The model’s verb class likelihoods for the individual semantic classes. 91

5.1 A visual representation of the two levels of abstraction in the model. 103
5.2 The difference between the log-likelihood values of the DO and PD frames. 110
5.3 Difference of log-likelihood values of the DO and PD frames over the course of training. 113
5.4 A visual representation of the two layers in the simulation of gradual Acquisition of Mental State Meaning. 117
5.5 The likelihoods of each event type given the learned constructions. 121
5.6 The distribution of our 5 verb classes over the 4 most frequent syntactic constructions for the MSVs. 121
5.7 The likelihoods of each event type given the learned verb classes. . . . . . . 123
Chapter 1

Introduction

The acquisition of Mental State Verbs (MSVs), such as think, know, and want, has been studied extensively by the psycholinguistic community, especially given the delayed onset of their production compared with other semantic verb classes, such as Action verbs, e.g., fall and go. Interestingly, the cognitive and linguistic properties of MSVs entail an array of challenges for the language learning child. First, MSVs describe inner mental states, such as beliefs and desires, that the learner must be able to conceptualize in order to identify them in a given context. Second, the learner must attend to the mental states in a given noisy language learning scene. That is, the child needs to focus on desires or beliefs that are not directly visible in the scene, as opposed to more visually salient alternative actions that may occur simultaneously, such as a running cat or a bouncing ball. Third, MSVs often refer to a related activity within a Sentential Complement (SC) construction, e.g., “She thinks daddy is making pancakes”, which requires the child to extract and map the semantic properties of two verbs appropriately. Thus, a comprehensive analysis of the MSV acquisition process is required to address the various cognitive and linguistic skills the learner must utilize while learning MSVs. This thesis analyzes the development of children’s ability to use MSVs to refer to mental states via a computational model that enables the representation of several of the above challenges and their interaction.
Psycholinguistic research has primarily focused on the challenge in conceptualizing mental states (e.g., Wellman and Estes (1987); Fodor (1992); Bartsch and Wellman (1995); Gopnik and Meltzoff (1997)), and this has been the focus of the computational analysis in this field as well (Berthiaume et al., 2008; Baker et al., 2011; Richardson et al., 2012). However, even when children can conceptualize belief and desire concepts, their ability to verbally refer to such concepts still lags behind (Denison et al., 2011; Rubio-Fernández and Geurts, 2012). Bartsch and Wellman (1995) argue that the delay in the production of MSVs beyond the point of conceptualization of mental concepts may result from the other challenges involved in referring to mental states in speech. Following such evidence, we focus here on this observed stage in which children have the ability to conceptualize mental concepts. We assume that at this stage in order to produce MSVs, children still need to learn to associate MSVs with the expression of mental meanings via grammatical linguistic structures. For example, a child may be able to conceptualize that people have inner belief states and yet have difficulty producing a grammatical utterance that refers to such a concept via a sentential complement as in “She thinks daddy is making pancakes”.

In this thesis, we assume that the semantic and syntactic properties of MSVs together with their distributional properties play a role in their acquisition. We represent a range of the properties of MSVs including their distribution over syntactic constructions, the various types of mental verbs (e.g., Belief verbs, such as think, and Desire verbs, such as want), etc. Specifically, we hypothesize that these linguistic properties of MSVs, and their distribution across MSVs and other verb classes, both hinder and facilitate the acquisition of MSVs as described below. We aim to analyze the acquisition of MSVs in light of the interaction of their semantic and syntactic properties across a variety of verbs and syntactic constructions.

We analyze our hypothesis by first analyzing the role of various semantic and syntactic
cues in delaying the acquisition of MSVs. We do so by evaluating how these cues may affecting the level of attention of children to the difficult-to-observe mental content of a scene and by that influencing the degree to which children can learn from such content. We then expand our analysis by broadening it to distributional properties of these cues across additional verb classes, such as Perception and Communication verbs. Our analysis supports claims about the facilitating role of the semantic and syntactic properties of MSVs that distinguish them as a verb class, e.g., the use of SCs and the reference to mental properties. Finally, we address additional aspects of the use of MSVs by evaluating the role of the variability of MSV usages across syntactic constructions, including non-SC constructions such as transitive and intransitive.

We follow a usage-based approach to language acquisition that allows us to test our hypothesis that patterns in the language input account for psycholinguistic observations of the acquisition of MSVs. In contrast to theories that suggest the existence of an innate mechanism to allow children’s rapid language learning process (Chomsky, 1981; Pinker, 1989), the usage-based perspective suggests that the language learning process can be facilitated solely by the information encapsulated in the language input children observe. We analyze the acquisition of MSVs using a probabilistic framework that can naturally capture gradient responses to the different cues in the linguistic input. The probabilistic framework of our computational model enables this analysis by modeling the learning of the association of verbs and relevant semantic and syntactic properties (Alishahi and Stevenson, 2008; Barak et al., 2012). Such a computational model captures the role of each linguistic property (e.g., the distribution of MSVs over syntactic structures) in the learning process while replicating acquisition patterns observed in children, e.g., the delayed production of MSVs in mental meaning. Moreover, by using an incremental model, we are able to investigate changes in behaviour over time given the naturalistic input that mimics the complexity of child-directed speech. Importantly, our model offers a novel approach to study the acquisition of MSVs that extends the psycholinguistic
literature on the topic. In addition, our model also extends previous computational models of verb learning that have mostly focused on concrete verbs that have visually accessible meanings, such as physical actions like go and eat (Alishahi and Stevenson, 2008; Chang, 2009; Perfors et al., 2010; Parisien and Stevenson, 2011). These verbs often occur with high frequency syntactic constructions with a single verb/clause, e.g., transitive and intransitive constructions. Chapter 2 reviews the psycholinguistic studies of MSVs and their acquisition, as well as relevant computational models. The following chapters present our computational experiments, their results, and their contributions as follows.

Chapter 3 presents our computational model that extends the model of Alishahi and Stevenson (2008) to address the linguistic and cognitive skills required to acquire MSVs.¹ The model of Alishahi and Stevenson (2008) is an incremental Bayesian model that learns argument structure constructions from naturalistic input (see Chapter 3 for a detailed description of the original model). We incorporate the representation of SC syntax into the framework of the existing model. In addition, we add a probabilistic component that simulates the difficulty in attending to the mental content, which is not directly observable from a visual scene, especially in the presence of a more salient physical action.

Our model captures the role of the interaction of the above additions to the model in the process of learning to associate MSVs to their mental meaning. Importantly, we show that both the use of SC syntax and the difficulty in attending to the mental content are required to replicate the observed difficulty of children and adults in attending to the mental meaning (Gleitman et al., 2005; Papafragou et al., 2007). This analysis broadens the perspective of the psycholinguistic study that looks at two age groups of children and adults (Papafragou et al., 2007). Importantly, the probabilistic framework of our model enables us to analyze the way children may interpret a usage of an MSV when they fail to attend to its mental content. Additionally, our probabilistic reasoning provides analytic

¹The work described in Chapter 3 was published in the paper by Barak et al. (2012).
tools to offer possible explanations for the psycholinguistic findings of Papafragou et al. (2007). We can estimate the likelihood of possible semantic interpretations of an MSV usage to shed light on the way a language learner might process such usages in naturalistic settings.

In Chapter 4 we use our model to investigate how the linguistic properties of MSVs may also assist children in learning to verbally refer to mental concepts. We study a key assumption of usage-based theories: that the acquisition of a construction relies heavily on the existence of a high-frequency exemplar verb that accounts for a large proportion of usages of that construction in the input (Casenhiser and Goldberg, 2005; Kidd et al., 2006, 2010). Importantly, unlike the psycholinguistic experiments that focus on the learning of an artificial novel construction using novel verbs, here we examine the acquisition of the SC construction from naturalistic input.

Our results provide new insights into exemplar-based learning in the context of naturalistic input with multiple semantic classes, and a diverse set of constructions for the verbs. Interestingly, the results replicate the observation of Kidd et al. (2006) that points to an early association of the finite-SC syntax to the Belief verb, think. This result emphasizes the role of the syntactic properties of MSVs, and Belief verb specifically, in their acquisition. The distribution of SC across verbs and verb classes leads to its early association with a Belief verb. However, at the same time, the use of SC may challenge its association with the proper semantic properties as shown in Chapter 3. Our framework leads to this evidence by enabling integrated analysis of the complex semantic and syntactic properties of MSVs.

We further our understanding of the trajectory by which our model learns to associate MSVs with the appropriate syntactic and semantic properties by comparing the strength of association of MSVs to particular semantic properties. The acquisition of Belief verbs lags behind the acquisition of Desire verbs in children. Some psycholinguistic theories

---

2The work described in Chapter 4 was published in the papers by Barak et al. (2013a,b).
attribute this lag to conceptual differences between the two classes, while others suggest that syntactic differences are responsible. Our model exhibits this pattern of acquisition, even though there is no difference in the model in the difficulty of the semantic or syntactic properties of Belief vs. Desire verbs. Our results point to the distributional properties of various verb classes as a potentially important, and heretofore unexplored, factor in the observed developmental lag of Belief verbs.

Our analysis thus far primarily focuses on the strong association of MSVs with the use of SC syntax. However, MSVs also occur in a variety of other syntactic structures (Diessel and Tomasello, 2001; Klainerman, 2010). A non-SC usage of an MSV may present fewer processing challenges to the language learner given a context with fewer arguments and a single verb (e.g., “I know his teacher” compared with “I know that his teacher gave him this book”). Moreover, other verb classes frequently occur with SCs, e.g., Communication and Perception verbs (de Villiers, 2005). The acquisition of the meaning of MSVs may be affected, by association, by such usages of an atypical syntactic property of MSVs (e.g., SC syntax) with a more readily visible meaning (e.g., saying). In Chapter 5 we present a novel extension to our computational model to learn verb classes, which allows us to analyze the association of mental verbs to their meaning given a higher level of generalization.\(^3\) Our incremental Bayesian model simultaneously and incrementally learns both argument structure constructions and verb classes given naturalistic language input. This incremental model offers a novel framework for analyzing monotonically growing verb classes as opposed to the previously-used batch algorithms to model learning of verb classes. Our results in this chapter point to an important role of the full syntactic preferences of MSVs, in addition to their occurrences with SCs. We show that our model initially interprets the meaning of MSVs as a non-mental meaning based on their usage with non-SC syntax such as “I want ice-cream” and “I know Chris”. This association to non-mental meaning provides experimental support to theories of early stages of MSV

\(^3\)The work described in Chapter 5 was published in the papers by Barak et al. (2014a,b).
acquisition in non-mental meaning, while considering both the well-studied use with SCs and the less-explored distribution over non-SC syntax.
Chapter 2

Acquiring Mental State Verbs

2.1 Overview

Mental state verbs (MSVs), such as think, know, and want, are produced by children after their third birthday, much later than less frequent verbs, such as run, see, and throw. Although some MSVs are of high-frequency, MSVs as a semantic group stand for only 10% of the verb usages in child-directed speech (CDS) (Howard et al., 2008). Nonetheless, studies that analyze the cause of the acquisition delay rarely point to the low frequency of MSVs in CDS as a determining factor for their acquisition. On the contrary, most psycholinguistic studies focus on the cognitive and linguistic properties of MSVs that impose challenges on the language learner (Bartsch and Wellman, 1995; Miller, 2006).

An MSV, such as think, refers to an inner state that is less salient than a physical action, such as sleeping that can be directly observed in a language learning scene. Moreover, an MSV is commonly used to express a relation to another event or state expressed within a sentential complement (SC) clause, e.g., “Sophia thinks he is sleeping”. In such cases, the learner needs to attend to the mental content, i.e., the thinking state, in the presence of the more salient physical action (sleeping event). At the same time, the
learner is required to extract the semantic relation between the two verbs as expressed by the syntactic structure, i.e., Sophia’s mental perspective on the *sleeping* action. As stated by Howard et al. (2008) among others, the delay in the acquisition of MSVs is suggested to be the result of an interplay of the development of the linguistic and cognitive skills associated with MSVs, rather than only one specific skill.

A computational model of the acquisition of MSVs can enable the analysis of the interplay among the various developing skills in order to understand their role in the observed developmental patterns of MSVs. Statistical computational models have been successfully applied to various aspects of verb learning, providing insights into the role of different syntactic and semantic cues in the acquisition of verbs and argument structure constructions. Recent modeling efforts have examined the possible role of learning generalization over the input verb usages, by forming an association between syntax and semantics at several levels (Alishahi and Stevenson, 2008; Perfors et al., 2010; Parisien and Stevenson, 2011). However, previous computational models do not directly address the linguistic properties associated with MSVs, and the cognitive development often considered as influencing the acquisition of these verbs. In particular, they do not directly consider complex syntactic structures (including a complement clause) that are central to the acquisition of MSVs. The following sections describe the challenges in acquiring MSVs, psycholinguistic theories of the acquisition of MSVs, and relevant computational models.

### 2.2 Challenges in the Acquisition of Mental State Verbs

Psycholinguistic theories have suggested that the acquisition of MSVs is delayed since certain cognitive and/or linguistic skills are not available during the early stages of language development. These skills include the ability to attend to a mental state in a
given language learning scene and utterance, the ability to represent the meaning of an
MSV, the ability to express such a meaning, e.g., one’s own desires or beliefs and also
those of others. When used in their genuine mental meaning, MSVs refer to the mental
states of the speaker or of others, e.g., “Daddy thinks this is a cookie, but it is a wheel”. Psycholinguists have suggested that very young children lack the cognitive ability to
conceive that others have mental states separate from their own (Bartsch and Wellman,
1995; Gopnik and Meltzoff, 1997), which in turn inhibits their ability to conceptualize
the meaning of MSVs.

The ability of young children to make inferences regarding the mental states of others
is mostly evaluated by their ability to report on a contradiction between the mental
state of an agent other than themselves and the state of the world. A false belief scene
is designed to enhance the existence of mental content by representing a contradiction
between the reality and a mental state, such as a belief state. For example, in the
most widely used false belief scene, also termed change-of-location scene (Fodor, 1992), a
puppet puts an object into a specific location, but then someone else changes the location
of the object in the absence of the puppet. In this case, the puppet is induced to have
a false belief about the presumed location of the object. The experimenter assesses the
ability of children to observe the puppet’s mental state via questions such as: “Where
does the puppet think the object is?”, or “Where will the puppet look for the object?”. 3-year-old children present below chance performance given such a task (Fodor, 1992;
Miller, 2006), while 4-year-olds consistently perform above chance (Fodor, 1992; Howard
et al., 2008), suggesting some developmental change within this time frame that facilitates
the change in performance.

An inability to perceive the mental states of others can explain the poor success rate
of 3-year-old children on false belief tasks, and also prevent children from using MSVs
in their genuine mental meaning at this developmental stage. However, false belief tasks
require the use of several cognitive and linguistic skills, and the lack of development
of each of these skills can result in the observed low success rate of 3-year-old children on such tasks. For example, Wellman and Bartsch (1988) demonstrate that 3-year-old children are able to understand beliefs as internal mental states to some extent, but have a specific difficulty in processing false beliefs. Fodor (1992) further narrows down the source of cognitive difficulty, by suggesting that 3-year-olds may be able to perceive false beliefs but that they are inclined to rely on their observations on the state of desires of others more strongly than on the beliefs of others, regardless of their truth value. The success in a change-of-location task also relies on the development of linguistic skills to understand verbal questions, such as “Where will the puppet look for the object”. In a modified version of a change-of-location task (Rubio-Fernández and Geurts, 2012), 3-year-old children were able to successfully act out the response of the puppet, instead of verbally describing it, when encouraged to play with the puppet and show “what is she going to do next?”. Any of the above developing cognitive and linguistic skills can offer a possible explanation for the lagging ability to pass standard false belief tasks, rather than an inability to perceive belief states altogether.

Several studies suggest that even when a child is aware that other people have mental states, MSVs are difficult to learn given their informational requirements (Gleitman et al., 2005; Papafragou et al., 2007; Howard et al., 2008). Children learn word meanings by figuring out which aspects of an observed scene are referred to by a particular word (Quine, 1960). Unlike physical action verbs, MSVs often refer to aspects of the world that are not directly observable (i.e., the beliefs and desires of another entity). In addition to the challenges posed by children’s developing linguistic and conceptual abilities, children may simply have difficulty in identifying the relevant mental content necessary for learning MSVs. This difficulty is enhanced by the frequent presence of other activities, given the frequent use of MSVs to refer to another activity, e.g., sleeping in “I think mommy is sleeping”. As suggested by Israel (2008), the child is then faced with a scene with a more salient activity, sleeping, while think should be mapped to the harder-to-observe
mental state.

Israel (2008) hypothesizes that the visual input from the world may impose another challenge on the language learner, since expressions of false belief may contradict visual evidence. For example, the utterance: “I think mommy is sleeping” may contradict the visual evidence of mommy being awake. To correctly represent the meaning of the verb *think* in such a usage, the child needs to override her own mental state of thinking mommy is awake, and represent a contrasting state of mind. Israel (2008) hypothesizes that the cognitive challenge in fully representing the meaning of MSVs relies on the required capacity for processing and integrating the various views of reality, rather than the ability to simply identify the existence of the mental state of others.

Nevertheless, the above mentioned difficulty is especially present when facing a false belief scene. There are other types of usages of MSVs that do not hold the same level of complexity. As Diessel and Tomasello (2001) point out, MSVs are often used in child-directed speech, not in their genuine mental meaning, but rather for conversational purposes to note that this is the point-of-view of the speaker. For example, a caregiver may say to a child, “I think it is time to go to sleep”, without expressing the uncertainty that is usually associated with the verb *think*. These usages, termed here as *performative usages*, require the child only to record that this is the perspective of the speaker. While these usages are not as demanding in cognitive requirements as the expression of false belief by a usage of an MSV, they may distract children from learning the genuine mental meaning of MSVs.

On top of the cognitive and semantic challenges described above, the complex syntactic structure that MSVs often appear with also imposes a developmental challenge on the language learner. MSVs typically occur with a sentential complement (SC) that refers to the propositional content of the mental state, as in “He thinks mom went home”. Children have to reach a stage of syntactic development that includes some facility with SCs in order to fully acquire MSVs. At 3–5 years old, children are able to process SCs
to some extent, but only imperfectly.

Notably, infinitivals appear earlier than finite SCs in the speech of young children (Bloom et al., 1984, 1989), though this acquisition pattern is in turn related to the strong association of each of the semantic subgroups of MSVs. In particular, Desire verbs occur mostly with an infinitival SC (as in “I want (her) to leave”), while Belief verbs occur mostly with a finite SC (a full tensed embedded clause, as in “I think (that) she left”). A wealth of research shows that children produce Desire verbs, such as want and wish, earlier than Belief verbs, such as think and know (Shatz et al., 1983; Bartsch and Wellman, 1995; Asplin, 2002; Perner et al., 2003; de Villiers, 2005; Papafragou et al., 2007; Pascual et al., 2008). Some explanations for this pattern posit that differences in the syntactic usages of Desire and Belief verbs underlie the observed developmental lag of the latter (de Villiers, 2005; Pascual et al., 2008). Others suggest that Desire verbs are conceptually simpler (Bartsch and Wellman, 1995) or pragmatically/communicatively more salient (Perner, 1988; Fodor, 1992; Perner et al., 2003). Proponents of the conceptual and pragmatic accounts argue that syntax alone cannot explain the delay in the acquisition of Belief verbs, because children use finite SCs with verbs of Communication (e.g., say) and Perception (e.g., see) long before they use them with Belief verbs (Bartsch and Wellman, 1995).

To summarize, the acquisition of MSVs requires conceptual and cognitive development, facility with the SC syntax, and the ability to attend to the cues of mental content. The development of these abilities occurs over the same period of time, and therefore their interaction over the course of acquisition of MSVs is particularly important. In addition, the distributional properties of the patterns in the input may also contribute to the process by which children acquire MSVs, such as the usage of MSVs with the SC syntax and the frequency of usage of MSVs in their genuine mental meaning. The psycholinguistic studies, described in the next section, analyze the possible interaction of these properties of MSVs as reflected in their acquisition.
2.3 Psycholinguistic Studies of MSV Acquisition

The first stage of MSV production, termed here as the performative stage, begins between 2–3 years of age and is characterized by using MSVs not in their mental meaning, but to direct the conversation, as an epistemic marker, e.g., “you are right, I guess”, or as an attention-getter, e.g., “know what?”. Moreover, while MSVs are identified with the usage of SC syntactic structure, the earlier performative usages of MSVs often appear in transitive or intransitive syntactic constructions, as in “I guess so” and “I don’t know”. This stage in the acquisition of MSVs has received much attention since the semantic and syntactic properties used for MSVs during this stage diverge from the properties associated with the true mental usages of these verbs. Studies of this stage first aim to define the properties of such usages as a separate type from the mental usages, and also to analyze their role in the acquisition process of MSVs (Shatz et al., 1983; Bartsch and Wellman, 1995; Diessel and Tomasello, 2001).

The second stage of MSV production, termed here as the mental production stage, begins only around 3 years of age, when children start using MSVs in reference to their genuine mental meaning. Even then, children still continue to develop their understanding of MSVs. Several studies, such as those by Moore et al. (1989) and Howard et al. (2008), show that the distinction between various belief verbs, such as think, guess, and believe, continues to develop until 8 years of age. Studies, such as Shatz et al. (1983), consider the transition from the performative stage to this stage as important evidence of the ability of children to perceive the mental states of others. The appreciation of the beliefs, thoughts, desires, and feelings of others, i.e., Theory of Mind (ToM), is required to successfully engage in daily communication, that is, to be able to motivate, reason, and anticipate the actions of others. Importantly, researchers, such as Miller (2006), observe the linguistic development from performative production to genuine mental usage as facilitating the development of Theory of Mind. We discuss studies of the performative stage and of the mental production stage in Section 2.3.1 and Section 2.3.2, respectively.


2.3.1 The Performative Stage

Recall that, the performative stage is characterized by non-mental usages of MSVs that serve various conversational functions, such as: repetitions of adult speech, epistemic markers, attention-getters, etc. (see examples for performative vs. mental usages in (1) and (2) below). Diessel and Tomasello (2001), for example, suggest that a type of non-mental usage that they term *formulaic* develops from the performative usages, and is limited to epistemic markers and attention getters. The formulaic usages are defined by Diessel and Tomasello (2001) as short phrases that serve a specific function in child-directed speech, such as using *I think* before another clause instead of the words *Maybe* or *Probably*. Since the distinction between subgroups of performative usages is not coherent across studies, we refer here to all non-mental usages as performative usages regardless of their semantic function or other linguistic properties, such as their syntactic properties.

(1) MSV usages marked as conversational (performative) usages by Diessel and Tomasello (2001) taken from CHILDES data

(a) “Know what?” (attention getter)
(b) “I think we should go” (politeness marker)
(c) “I think so/not” (yes/no reply)
(d) “You are right, I guess” (epistemic marker)
(e) “I don’t know” (expression of refusal)

(2) MSV usages marked as genuine mental usages by Bartsch and Wellman (1995) taken from the CHILDES datasets (Brown, 1973)

(a) “Because I want it to ride in de back. Why do that? Do you think it’s going to stay on dere?” (Use of question and reference to mental state of other)
(b) “I don’t know. He ran down the street?” (Marked as genuine reference based on the context, a reply to a wh-question)
(c) “You thought that was the wheel to that” (Use of contrastive, a contradiction to the reality described in the conversation)

The performative stage consistently appears as the first stage of production of MSVs across children and verbs (Diessel and Tomasello, 2001). Howard et al. (2008) and Israel (2008) suggest that this is a required stage in the acquisition of MSVs since the performative usages are easier for children to process given their simplified syntactic and semantic requirements. For instance, Howard et al. (2008) provide evidence for single-clause MSV utterances in CDS as being predictive of the later production of MSVs by children. They suggest that the simpler syntactic structure enables children to focus on the mental content rather than the processing of the syntactic information. However, since performative usages express a semantic function other than mental meaning, their contribution to the acquisition of genuine mental meanings of MSVs may be limited.

A leading perspective on the role of performative usages in the acquisition of MSVs argues that performative usages are an unrelated usage of MSVs that do not share semantic or even syntactic properties with the mental usage of MSVs. For instance, Diessel and Tomasello (2001) propose that the performative stage may not have a specific role in the process of learning MSVs, but simply reflect the frequency of performative usages in CDS. While Diessel and Tomasello (2001) base this proposal on the high frequency of performative usages in adult-to-adult speech (Thompson, 2002), evidence from CDS imply that the performative stage is unlikely to be simply the result of mirroring adult-speech (Furrow et al., 1992; Moore et al., 1994; Bartsch and Wellman, 1995). According to Bartsch and Wellman (1995), the average rate of performative usages of the verbs think and know in child-directed speech reaches a rate of about 30%, much lower than the rate reported by Thompson (2002) for adult-to-adult speech. The pronounced differences across these two studies could be attributed to the differences in properties of CDS and adult-to-adult speech, but also to the annotation method of studies which focused more on semantic properties of the usage in CDS (Furrow et al., 1992; Moore et al.,
1994; Bartsch and Wellman, 1995) rather than the syntactic and lexical properties as in Thompson (2002). By focusing on the semantic properties of MSV usages, Furrow et al. (1992) and Moore et al. (1994) present a difference in the distribution of communication goals of children and adults in their use of the verbs think and know. They find that parents use know with higher frequency than think and more often to refer to mental states. On the other hand, they observe that children use think rather than know with higher frequency, mostly for non-mental usages. The differences in the semantic goals of the usages of MSVs across parents and children in these studies imply that the production of performative usages by children do not simply reflect their distribution in CDS.

A different interpretation of performative usages argues that in such usages the head clause containing the MSV actually functions linguistically as a subordinate clause given its semantic properties (Thompson, 2002; Kearns, 2007; Van Bogaert, 2010). That is, the head clause of these MSV usages is similar in meaning to an adverbial that expresses a stance, such as probably or surely. As illustrated by Diessel and Tomasello (2001), the head clause in the utterance: “I believe that this is a mistake”, can be replaced in the following way and maintain the original meaning of the utterance: “This is surely a mistake”. These studies question the ability to interpret performative usages as syntactic patterns with a subordinate clause.

Assuming performative usages do not truly refer to two actions, Diessel and Tomasello (2001) offer that the performative stage is a result of limitations in the computational power of children. The computational power presumably limits children to producing utterances that contain a single verb and prevents them from using SCs and propositional relations to express the true mental meaning of MSVs. That is, children would be able to produce, under the computational power restriction, utterances such as “I don’t know”, but also “I think that’s a car”, when “I think” is considered as a formula or a phrase rather than as a pronoun followed by a verb. Bartsch and Wellman (1995) argue against the computational power hypothesis by bringing evidence from the use of other verbs,
such as see and say, with SCs before the appearance of the performative stage with belief verbs.

An alternative perspective suggests that the performative usages represent a partial semantic meaning to the full-mental meaning that MSVs express. For instance, the usage of MSVs across stages can be viewed as a changing rate in the level of certainty expressed by MSVs (Howard et al., 2008). For verbs such as think and guess, the performative usages express a higher level of certainty than their meaning when referring to mental states. For example, the verb think is used in (1)b as a politeness marker while the speaker has high certainty level that “We should go”. Several studies, such as those by Nixon (2005), Pascual et al. (2008), and Howard et al. (2008), present empirical evidence of the production of high certainty usages of MSVs by children. These studies observe a strong association of MSVs to the expression of certainty in naturalistic child speech during the initial production of MSVs. Although these usages are not annotated as performative usages, the initial stage of MSV usages expressing certainty corresponds to the performative stage marked by other studies discussed above, and may serve as an alternative interpretation of this stage.

Performative usages are often focused on expressing the level of certainty of the speaker. Usages that express certainty simplify the challenge of inferring the meaning of MSVs, since these usages encode less information in addition to the embedded event, compared with an expression of genuine mental meaning. Such usages are less likely to contradict the observable reality, e.g., “I think giraffes are tall” is unlikely to be used in contradiction to the reality. The language learner is required to map this use of think to the speaker expressing a personal point-of-view, but is not required to make a complex integration with a contradicting reality. When the MSV is used to express genuine mental meaning in an utterance with an SC, studies, such as Israel (2008) and Miller (2006), suggest that the semantic relation between the two clauses is the main source of the observed acquisition challenge.
The mental usages of an MSV mostly do not focus on the certainty of the speaker. By encapsulating the speaker’s mental state in regards to the embedded event in the SC, genuine mental references encode additional, and possibly complex, information further to the action in the SC. For example, a language learner must learn the semantic distinction between the verbs *think* and *guess* when used with uncertainty, since they mark different levels of belief. False belief usages mark, according to Israel (2008), the highest level of semantic complexity since they require incorporating two contradicting perspectives of the reality. Israel (2008) hypothesizes that children are cognitively unable to hold the numerous point-of-views on reality at the same time. However, this hypothesis does not apply to the performative usages, which therefore are accessible in earlier stages to serve as an initial representation of the meaning of MSVs. Israel (2008) claims that children go through cognitive development that allows them to eventually process several points of view and even contradicting ones. This is in contradiction to Bartsch and Wellman (1995) and de Villiers (2005), who argue for a difficulty in the perception of the mental state rather than its processing or representation.

Although the semantic and syntactic properties of genuine mental usages of MSVs may impose difficulties on the language learner, these properties may also enable the transition from the performative stage to the stage of mental production of MSVs. The contradiction between the mental point-of-view and reality (as in a false belief situation) has been shown to enhance attention to the mental content in a scene (Papafragou et al., 2007; Pham et al., 2012). Similarly, the use of complex syntax (i.e., SCs) increases the ability of children and adults to identify MSVs in a verb identification task (Gleitman et al., 2005; Papafragou et al., 2007). The association of MSVs with the SC syntax may highlight the reference to mental meaning in typical language learning scenes similar to the experimental settings of Papafragou et al. (2007). The following section discusses the linguistic cues that may assist the acquisition of MSVs in their true mental meaning.
2.3.2 The Mental Production Stage

The use of MSVs to express genuine mental meaning marks the reference to a more complex meaning than the meaning of MSVs in performative usages. This complexity challenges the language learner in identifying the real-world referent of the MSV much more than for verbs with a higher level of concreteness, such as a visual communication act expressed by the verb *say*. Even for polysemous non-mental verbs, such as *see*, the acquisition of the concrete physical meaning has been suggested to precede the acquisition of the verb’s abstract meaning, as in “I see your point” (Johnson, 1999). Gleitman et al. (2005) make a more general claim regarding the order of acquisition of words in regards to their level of concreteness. Their analysis of child productions reveals that, e.g., the verb *kiss* appears in child language before the noun *idea*, even though young children generally produce nouns before verbs. Gleitman et al. (2005) hypothesize that more abstract words are harder to learn via a word-to-world mapping procedure, where children learn mappings between a noun and an object, or a verb and an event/state. They suggest that for verbs with less salient semantic content, the syntactic structure holds an even more important role in the acquisition process than for verbs that refer to concrete physical actions.

Verbs that share semantic properties often appear in similar syntactic constructions (Levin, 1993). Based on this distributional property, Gleitman (1990) suggests a process of *Syntactic Bootstrapping* by which children can acquire the semantic properties of a verb based on its distribution over syntactic structures. As a first step, a child can map the components of a syntactic structure (e.g., *subject*) to their semantic roles (e.g., *agent*) based on a verb (e.g., *see*). Following this step, the child can generalize the extracted semantic relation between the arguments of the structure to a new verb (e.g., from *see* to *think*). de Villiers (2005) presents such a bootstrapping theory of the acquisition of MSVs based on their syntactic properties. According to this theory, as a first stage children associate MSVs to more concrete verbs – such as the communication verb *say* – because
of the shared SC syntactic structure. At this stage, as she hypothesizes, children can develop the ability to express a point-of-view as a shared component of the meaning of communication and belief verbs. For instance, both “I said mommy is sleeping” and “I think mommy is sleeping” express the perspective of the speaker using the same syntactic structure, which is frequent with both communication and belief verbs. Bartsch and Wellman (1995) present evidence of earlier usages of communication verbs with finite SC compared to the production of finite SC with belief verbs. This empirical evidence from the acquisition process of SCs supports the hypothesis of de Villiers (2005) who argues that an earlier acquisition of communication verbs is required for the acquisition of belief verbs.

As a second step in the theory of de Villiers (2005), the language learner can attend to the semantic properties that differentiate MSVs from communication verbs i.e., their reference to mental content, and the possibility of contradicting reality. de Villiers (2005) argues that finite SC utterances are unique in their ability to represent a true statement even when the SC has a negative truth value. For example, the statement “I said/think mommy is sleeping” above can be true even when mommy is awake, regardless of the choice of verb. Finite SCs differ from infinitival SCs in this property, as infinitivals can refer to an unfulfilled desire but not to a proposition that directly contradicts reality. That is, “I want mommy to sleep” refers to a future activity of mommy without contradicting a current observation.

de Villiers (2005) hypothesizes that children move on to a second stage in the acquisition of the mental meaning by associating MSV usages to the ability of finite SC to hold a true value while contradicting reality, by using usages of communication verbs in a similar context. Children can observe a use of an utterance such as “he says mommy is sleeping”, while they observe that mommy is awake, as the saying action is directly perceivable from the scene. de Villiers (2005) suggests that the association of such an utterance to another utterance that uses an MSV, e.g. think, can facilitate the association
of MSVs with the ability to express false-belief.

Empirical evidence for the underlying role of the syntactic structure in the comprehension and detection of MSVs is evident in the results of experimental studies with adult participants. To investigate the linguistic and contextual cues that could help in learning MSVs, Gleitman et al. (2005) use a procedure called the Human Simulation Paradigm, originally proposed by Gillette et al. (1999). In this paradigm, subjects are put in situations intended to simulate various word learning conditions of young children. For instance, in one condition, adults watch silent videos of caregivers interacting with children, and are asked to predict the verb uttered by the caregiver. In another condition, subjects hear a sentence containing a nonce verb (e.g., gorp) after watching a silent video, and are asked what gorp might mean. The results of a study by Gleitman et al. (2005) show that participants predict mostly physical action verbs based on the silent video input, while an utterance with an SC significantly increases the rate of predicting MSVs. Papafragou et al. (2007) confirm this finding for children as well as adults in a similar novel verb prediction task.

Following the results of Gleitman et al. (2005), Papafragou et al. (2007) use the Human Simulation Paradigm to evaluate the role of SC and false belief cues in an MSV prediction task performed by children and adults. The contribution of contrastive semantic input to this task was examined using a video that described either a True Belief or a False Belief scene: A True Belief scene shows an ordinary situation that unfolds as the character in the scene expects, e.g., a little boy takes food to his grandmother, and she is there in the house as expected. The corresponding False Belief scene has an unexpected outcome for the character, e.g., in the above example another character has replaced the grandmother in her bed.

Papafragou et al. (2007) show that even given adequate conceptual and linguistic abilities (as in adults), the mental events in a scene (the actors’ internal states) are not attended to as much as the actions, unless there are cues - as in the false belief condition -
that heighten the salience of the mental content. Moreover, their results show that adults
and children are sensitive to both the semantic and syntactic cues, and that their effects
add up when they are presented together. However, children’s ability to draw on such
cues is inferior to that of adults. Papafragou et al. (2007) thus propose that the difference
between children and adults is that children have not yet formed as strong an association
as adults between the cues and the mental content of a scene as required to match the
performance of adults. Nonetheless, their results suggest that the participating children
with a mean age of 4;5 had the conceptual and linguistic abilities required for identifying
MSVs, since they were able to produce them under conditions with sufficiently strong
cues.

Another source of evidence of the ability of children to associate MSVs to SCs at a
similar developmental stage comes from the analysis of their performance on an imitation
task with finite SCs. In this task, children are asked to repeat an utterance word-by-
word. The performance on the task is evaluated with respect to the linguistic properties
of the utterance. For instance, the evaluation can measure the correlation of the success
rate with the use of SC with low and high-frequency MSVs. Kidd et al. (2010) asked
children to repeat a given utterance with a finite SC word-by-word, while controlling
the frequency of the head verb with finite SCs. Their results show that when children
mistakenly substitute the uttered verb with a verb of their choice, they choose the MSV
think 88% of the time, regardless of the semantic class of the given head verb. Kidd et al.
(2010) suggest that the strong association of finite SC to the verb think is the result of
its high frequency of appearance with this construction. Moreover, following Goldberg
et al. (2004), they hypothesize that think functions as a prototypical verb for SCs. That
is, think would facilitate the acquisition of the finite SC structure by associating this
structure with the ability to express mental meaning.

Such observations from psycholinguistic experiments raise the question of what leads
children to limited production of MSVs before the age of 4 years old, and how this affects
their success in ToM tasks that test their ability to refer to mental states of others. Such evidence calls for the analysis of the interplay of the various semantic and syntactic cues, and also of the interplay of the children’s developing linguistic and cognitive abilities. As discussed in the following section, computational models facilitate such analyses by enabling a focused investigation of the various settings of the input, as well as the learning mechanisms the computational learner requires to present similar patterns of acquisition of verbs in general and of MSVs in particular.

2.4 Computational Models

The acquisition of verb meanings relies on linguistic properties, such as syntactic structure and lexical properties, much more than nouns that often can be directly mapped to an object that is referred to. The meaning of a verb captures how the verb connects nouns (associated with entities). For instance, the verb *drop* in the utterance “the child dropped the ball” expresses a causative semantic connection between *the child* and *the ball*.

When exposed to an utterance containing a verb, a child may use the meaning of the nouns in the utterance, and of the syntactic construction, to determine the meaning of the verb. Psycholinguistic studies show that children are sensitive to the distributional properties of CDS, such as the distribution of verbs with similar syntactic structures (Gillette et al., 1999), in the process of verb acquisition. It is debatable to what extent children rely on innate language abilities, as suggested by Chomsky, and to what extent the statistical properties of the language result in the various observed phenomena in the way children begin to use a language. Computational models of verb learning can simulate a learner, while explicitly defining the prior linguistic knowledge of the learner, the learning capabilities, and what the learner receives as input. These models can be particularly interesting when providing information on which of the assumed capabilities and knowledge sources produce developmental patterns that are similar to those observed
in children.

Computational models enable the analysis of theoretical hypotheses regarding various aspects of the language acquisition process. For example, computational models have been successfully used to analyze specific phenomena in language developments, such as word learning (Siskind, 1996; Fazly et al., 2010), verb argument structure acquisition (Alishahi and Stevenson, 2008; Perfors et al., 2010; Parisien and Stevenson, 2011), and acquisition of syntactic structures (Freudenthal et al., 2007). Such models enable the processing of large amounts of language input, currently available through the CHILDES database (MacWhinney, 2000), which includes a large amount of child-directed and child language across a variety of age ranges. By replicating a language acquisition pattern in a computational learner, the computational evidence illustrates how such patterns can be the result of the distributional properties of the language and statistical inferences from them. Moreover, such a model explicitly defines its assumptions on the required inference abilities of the learner and previously acquired knowledge, and can show the dependency between the assumptions and the perceived patterns. For instance, such a model can show how children may learn that the use of verbs in certain constructions, such as “I fell the ball”, are ungrammatical by making statistical inferences over CDS input, i.e. without making an assumption about the availability of any explicit negative feedback (Alishahi and Stevenson, 2008). Such evidence from computational models supports theories of statistical learning, by showing the ability to make inferences from input that holds the same characteristics of CDS.¹

Computational models have also been used to study the cognitive development of children as it is reflected in their ability to make inferences regarding mental states of others, i.e. development of Theory of Mind (ToM). The ability of children to address genuine mental states of others in their productive usage of MSVs brings evidence to

¹We focus here on computational models that use verb clustering to model verb acquisition rather than models that aim at using distribution of semantic and syntactic properties of verbs to create coherent verb clusters, or even a hierarchy of meaningful verb clusters. For such models please refer to (Schulte im Walde, 2008; Joanis et al., 2008; Sun and Korhonen, 2011).
Chapter 2. Acquiring Mental State Verbs

their ability to make inferences regarding such mental states. For this reason, computational models of ToM provide us with an experimental framework for modeling cognitive development related to the acquisition of MSVs. We first review models of ToM, which analyze the development of cognitive mechanisms that may facilitate the acquisition of MSVs. We then review computational models of verb acquisition and the acquisition of grammatical constructions, which refer to language development processes relevant to the acquisition process of MSVs.

2.4.1 Computational Models of Theory of Mind

Theory of Mind (ToM) refers to the ability of children to identify other people as having a state of mind, i.e., inner states such as desires, emotions, and beliefs that are different from that of the child. Psycholinguistic studies of ToM have often focused on the ability of children to make references to the observation of false-belief of others (see Section 2.2). Specifically, these studies have analyzed the performance of children on false belief tasks under the assumption that some developmental changes facilitate the eventual success of children on such tasks. Computational models of ToM aim to simulate the developing cognitive skills of children, which facilitate their success on false belief tasks. Such models clearly define the available input and the inference abilities of the computational learner to present how a modification in either the input to the model or its modeled inferences may lead to an improvement in the success rate in predicting the belief state of others, i.e., a false belief task.

Berthiaume et al. (2008), for instance, offer a connectionist model to emulate the developing performance of children in a false belief task. The input to their network represents the information available to a child in a typical false belief task, i.e., the first location of the object, the second location of the object, whether the puppet was present when the object changed location, and an additional variable to allow stochastic learning. “Younger children” are represented by a network with a lower number of
hidden units compared with the computational “older children”. They argue that the
novelty of the paper lies in that the cognitive development is modelled as a change in the
inference power, i.e., the number of hidden units, rather than by changing the parameter
settings of the model manually as required by previous models, such as those by Triona
et al. (2002). In this model, the performance of younger children requires decreasing the
number of hidden units, which may mean a difference in the processing of the information
despite its availability. In that, the authors make an assumption on an increase in the
cognitive capacity of children at different stages of ToM development, without specifying
the nature of this cognitive change.

Baker et al. (2011) create a richer model of the world observed by the computational
learner, by representing the actions and desires of the agent, typically represented by a
puppet in psycholinguistic studies, on top of the agent’s belief. In this Bayesian model,
the learner faces a challenge of predicting the actions of the agent while combining an
estimation of both the desires and beliefs of the agent, rather than the typical false belief
scenario, which refers only to the belief state. The learner of Baker et al. (2011) performs
the best when considering both beliefs and desires, suggesting that psycholinguistic stud-
ies of ToM should consider both types of mental states in the settings of a belief study.
Richardson et al. (2012) compare the performance of their model to the inference abilities
of children in a psycholinguistic experiment with similar settings. Specifically, Richard-
son et al. (2012) evaluate the information needed within the model, i.e., desires, beliefs,
or both, to resemble the performance of children at different developmental stages in the
age range of 3 to 6 years. Their model mimics the behaviour of younger children when it
focuses on desires, and the ability of older children when it shifts focus to beliefs, similar
to the theory presented by Fodor (1992) (see Section 2.2).

Unfortunately, none of these models takes into consideration the mental state of the
learner itself. That is, the models incorporate the beliefs, and sometimes the desires,
of the observed agent, but do not capture the way the learner integrates those states
with its own inner state or resolves possible conflicts. For example, a child may have personal beliefs regarding the location of the object, or personal desires regarding which object the agent should find, and these beliefs and desires may be in conflict with those of the agent. In contrast to that, computational models of ToM mostly manipulate the perceptual ability of the learner, rather than allowing the learner to gradually improve in its ability to hold various types of information from various agents including the learner itself.

Regardless of these shortcomings, these models address the cognitive development of children that may limit their competence in making inferences about the mental states of others. In turn, the limited success in making inferences is likely to influence the ability of a language learner to identify the meaning of MSVs in genuine mental usages given a typical language learning scene. Such developing cognitive abilities have not been considered in the context of computational models of verb learning or construction learning, which are reviewed in the next section. On the other hand, none of the computational models of ToM consider the linguistic requirements of standard false belief tasks that often include answering verbal questions. In this sense, computational models of ToM and computational models of verb learning are complementary to each other from the perspective of acquisition of mental verbs.

### 2.4.2 Computational Models of verb learning

Initial production of verbs follows that of nouns, possibly because the meaning of a verb relies on the way it connects the nouns in a sentence. For example, the verb *drop* in “I dropped the ball” expresses a semantic relation between the nouns *I* (agent) and *ball* (theme). However, once children are able to associate a syntactic structure (e.g., “arg1 Verb arg2”) to a certain meaning, they can generalize the meaning (e.g., an agent performing an action on a theme) of the acquired verb to new verbs that appear in the same syntactic structure. Children have been shown to demonstrate this
ability to generalize by correctly predicting the appropriate meaning for a novel verb, e.g. *gorp*, appearing in a familiar syntactic structure, e.g. “The child gorped the cup”. The prediction of the meaning can be demonstrated, for example, by picking one of two possible animated semantic alternatives, or by offering a possible verb that can be used instead of *gorp*. Computational models have been found useful in the analysis of such generalization abilities. A computational model can be used to understand whether and how statistical regularities in CDS can lead to stages in the generalization process that correspond to similar stages observed in child productions. For example, Alishahi and Pyykkonen (2011) show an improved ability to correctly predict a novel verb when a computational model also relies on syntactic information on top of semantic information.

Earlier models of verb learning have utilized a variety of computational frameworks to analyze the importance of the association of syntax and semantics in verb learning; however, these models have been limited in the range of analyzed syntactic constructions. For example, Dell’Orletta et al. (2005) model the role of high-frequency verbs in allowing children to correctly interpret “arg1 verb” as either a Subject-Verb or Object-Verb construction. Their model offers initial evidence on how the statistical regularities of linguistic features, such as the animacy of a noun, can facilitate the association of argument structure constructions with specific groups of verbs. Despite the limitations of their data in represented constructions and range of features, their model is able to acquire a robust syntactic generalization over this pattern.

Desai (2002) utilizes a richer representation of the input in his studies of the association of syntactic and semantic cues in the process of acquiring verbs, compared with Dell’Orletta et al. (2005). This connectionist model simulates the association of transitive and intransitive constructions to their meaning given a simple representation of the scene as either action, e.g., “a boy is jumping”, or causative, e.g., “a girl is pushing a boy”. The model holds both the ability to bootstrap from the syntactic structure to identify the meaning of a novel verb, and also shows how the meaning of familiar nouns
improves the ability to learn those syntactic structures. However, the input to the model still addresses a narrow range of only two constructions. The model is only evaluated on a small artificial language that does not resemble CDS in the complexity of constructions, nor richness of nouns and verbs.

The model of Niyogi (2002) encodes more naturalistic semantic and syntactic information, although it is also simplified within a small artificial language. The model requires only a small number of input items to predict the meaning of novel verbs, though it relies on manually annotated data and built-in knowledge on the distribution of the verbs in the input over the constructions. Another example of how semantics and syntax can be used to learn the meaning of novel verbs is offered by Chang (2004). This model aims to learn constructions by associating motion verbs, such as throw and fall, with patterns of usage with syntactic structures and semantic properties. The model learns verb-islands, i.e., each verb forms a construction of its own initially (Tomasello, 2003). The construction enables the model to generalize over the semantic properties of the nouns used with each of the verbs, e.g., based on specific usages with throw, the model learns the subject of throw is typically an animated agent. However, this model is still extremely limited in the included constructions, and requires extensive manual annotation of the semantic representation of the input. Moreover, the generalizations over the verb usages are limited compared with the generalizations presented by Desai (2002) and Niyogi (2002).

More recent models of verb learning have relied on more naturalistic representations of the input. Freudenthal et al. (2007) utilize a graph-based algorithm that models the acquisition of grammatical structures. Their model records specific verb usages and allows generalization across them by connecting verbs that co-occur with the same nouns. The simplicity of their framework enables the use of natural CDS without requiring extensive annotations. Moreover, this framework allows the representation of various syntactic structures, including multiple-verb utterances. They achieve this by representing each word independently in a graph without encoding any syntactic or semantic knowledge on
the words or the utterances. Despite the advantage of this simplistic representation it also results in a lack of any semantic representation that is available to children while learning a language, and that is especially important for the acquisition of verb meaning. Despite the flexibility of the framework, the learning algorithm builds in a specific learning bias of greater attention to words appearing in utterance-final positions. This property of the model is crucial for its ability to replicate the tendency of children to produce partial utterances which focus on only the SC, which is the second clause. Although this work does not focus on the acquisition or production of MSVs, this finding is inline with the performatative stage of MSVs described in Section 2.3 that is characterized by an earlier production of MSVs as an epistemic marker while focusing on the SC. Nevertheless, this is one of the only models that analyzes the acquisition of syntactic structures having several verbs, e.g., utterances with sentential complements.

Fleischman and Roy (2005) focus on the semantic representation of an utterance, rather than the syntactic development. The novelty of this study is the analysis of verbs given their ambiguity as intentional actions meant to lead to some desired outcome. The authors model two types of ambiguity of an observable action: the level of description of the action, and the intended action behind the observable act. For instance, when an agent opens the door of a house, a low level description would refer to the open action; however, in a higher level description the action could be interpreted as leave, as in “he left the house”. The second type of ambiguity refers to the possible alternative interpretations of an action as a preliminary action intended to facilitate an additional action, i.e., open can be the first action in order to leave the house, but also to let in, as in “he let the guest in the house”. The Expectation Maximization algorithm is used to find the most likely interpretation of all nouns and verbs in an utterance, by considering all possible interpretations out of a predefined hierarchy of intentional actions that addresses the two types of ambiguity. In each step, the model considers all actions in the scene as a possible mapping to the verb and nouns in the utterance, and
also considers other interpretations of the action based on the agent’s goal. Importantly, this model performs much better on nouns than on verbs mostly due to the sparseness of verbs in the hierarchy. In addition, the model requires extensive annotation in order to build the semantic hierarchy of intentional actions.

Connor et al. (2012) model use a unified framework that learns semantic role labels and syntactic representation. Their unsupervised approach relies on plausible background knowledge and realistically noisy semantic feedback as opposed to other systems that require full annotation of the semantic properties of the scene, as in Fleischman and Roy (2005) above. Their model extracts properties of verb usages that are directly evident from the utterance, such as, the position of arguments relative to the verb, to infer the semantic roles of the arguments. Their method can capture differences across transitive and intransitive usages of the syntactic preferences of verbs as in “Chris and Sarah ran” and “Bill came” vs. verbs that require an object, as in “I sent the letter” and “She dropped the ball”. The model mimics the behaviour of children by identifying the input “Chris and Sarah ran” mistakenly with Chris as the agent and Sarah as the theme rather than a second agent. However, the model is limited in its ability to represent the fine-grained event properties that can distinguish the meaning of MSVs from other semantic classes, e.g., Perception verbs.

The Hierarchical Bayesian model of Kemp et al. (2007) has been used in several studies to learn verbs and constructions (Perfors et al., 2010; Havasi and Speer, 2011). Kemp et al. (2007) first applied their model to learn the use of semantic regularities in the process of word learning, however the model can incorporate any probabilistic feature regardless of its linguistic role, e.g., semantic, syntactic, or lexical. The model observes input instances in its first layer, and generalizes these observations to probabilistic parameters over the occurrences of the input instances or the grouping of the instances, i.e., it learns over-hypotheses. The model is flexible in the number of layers of over-hypotheses it learns, in the type of over-hypotheses it uses, and in the number of clusters it creates.
Havasi and Speer (2011), for instance, use this model to analyze biases in the association of motion and path verbs, such as *go* and *exit*, to one of two syntactic constructions in the process of their acquisition. The goal of Havasi and Speer in this study is only to analyze the ability of a language learner to correctly identify a novel verb as either a motion verb or a path verb based on a given syntactic pattern, e.g., to recognize that the novel verb in “She is going to *rapple* across the road” is more likely to be a motion verb rather than a path verb. Based on the results from an earlier study, Havasi and Speer (2011) use a multinomial distribution to learn the association of each group of verbs to its typical grammatical structure. However, the authors argue that generalization over the input is required in order to enable correct induction of these grammatical preferences to novel verbs. Therefore, they utilize the model of Kemp et al. (2007), but assume the association of each verb from the input to its meaning has been learned, and only aim to generalize this knowledge to new verbs, i.e., the semantic description of the verbs encodes prior knowledge about their association to the appropriate class. Although this application of the model uses an artificial semantic representation that only specifies the distinction in question (i.e., motion vs. path), Havasi and Speer illustrate the flexibility of the model and its ability to correctly learn form–meaning associations.

Perfors et al. (2010) present another application of the model of Kemp et al. (2007) to the acquisition of verb argument constructions. Similar to Havasi and Speer (2011), Perfors et al. (2010) assume that each verb usage is associated with an appropriate semantic meaning and syntactic construction. However, Perfors et al. (2010) extend the original model with a higher layer of abstraction that learns classes of verbs that represent the association of verbs over verb alternations, i.e., groups of verbs with similar patterns of occurrences across syntactic patterns. For example, a verb such as *give* that occurs in the prepositional dative, e.g., “John gave the book to Heather”, often occurs also in
the direct object dative, e.g., "John gave Heather the book". By incorporating verb classes into the model, their computational learner is able to mimic the human ability to predict the correct syntactic patterns for novel verbs, i.e., to predict, after exposure to the new verb in one of the syntactic structures, its ability to occur in the other. The model also learns to restrict the generalizations over the verb alternation, to prevent the generalizations to ungrammatical utterances by applying the dative alternation on verbs that cannot take this pattern, e.g., confess in "* John confessed Heather the truth". However, the model is assessed on a simplified alternation behaviour of the verbs. The data used for assessment only encodes two simple syntactic patterns, and with only three semantic features that correspond to the semantic distinction across the patterns. By that the model assumes the prior availability of the association of meaning to a syntactic construction, and also the prior learning of the constructions themselves. Despite the ability to mimic child behaviour in verb learning by Perfors et al. (2010) and Havasi and Speer (2011), their models simplify the learning process by relying on the knowledge of constructions. Moreover, although the input to their models uses natural CDS, the input is limited to these two constructions and the verbs that appear with them. As a result, the input contains a much reduced number of verbs, constructions, and alternations than naturally occur in CDS.

Parisien and Stevenson (2011) address some of the limitations of the model of Perfors et al. (2010) in a Hierarchical Dirichlet Model based on the model of Teh et al. (2006). The model of Parisien and Stevenson (2011) incorporates a layer for the representation of the verb classes based on verb alternations, but also a lower layer that learns the verb constructions, which Perfors et al. (2010) assume as prior knowledge. Moreover, their model incorporates naturalistic features of semantic, syntactic, and lexical properties of the verbs, and does not limit the input to only the dative constructions, as in Perfors et al. (2010). The learner of Parisien and Stevenson (2011) addresses generalization questions regarding the dative alternation as in Perfors et al. (2010), and is able to mimic child
behaviour given the challenge of more naturalistic data. Although the data for the model includes a variety of verbs and constructions, the model is not evaluated on utterances with multiple verbs, i.e., constructions with SCs.

The framework of the Incremental Bayesian Model presented by Alishahi and Stevenson (2008), also addresses some of the limitations of the model of Perfors et al. (2010), by modeling the acquisition of constructions. The model of Alishahi and Stevenson (2008) learns verb constructions as the association of semantic and syntactic properties over verb usages. The model enables the prediction of the semantic properties of a novel verb, e.g., *gorp* in “he gorp the ball”, given the learned associations over the construction, or alternatively the prediction of the syntactic pattern from the semantic properties. Since the model implements incremental learning, it enables the simulation of a developing facilitation with constructions similar to child language acquisition. For example, the model presents stages of possible over-generalizations to predict the use of the verb *fall* in the transitive construction, e.g., “He fell the ball”, and also the recovery from such erroneous stages. This pattern of language learning is achieved by generalizing over the learned constructions, and over time recognizing the lower likelihood of some verb to be associated with a certain construction, such as *fall* with the transitive construction. However, the model is limited in the range of simulated learning patterns as it does not model the learning of higher level generalizations over the constructions, such as verb alternations, during the process of learning. Moreover, the incremental process of the model, although beneficial for simulation of developing abilities, makes it more sensitive to input order and noise. Importantly, the semantic representation of Alishahi and Stevenson (2008) is based on an assumption that the scene contains a single semantic interpretation of the action, which is a simplification of a typical language learning scene that includes several actions carried on in parallel.
Chapter 3

Attending to Mental State Verbs

3.1 Introduction

The cognitive requirement of MSVs, i.e., the ability to conceptualize and identify inner mental states, has been at the center of research regarding the acquisition of MSVs. Psycholinguists have suggested that very young children lack the conceptual ability to conceive that others have mental states separate from their own (Bartsch and Wellman, 1995; Gopnik and Meltzoff, 1997), further delaying acquisition of MSVs beyond the acquisition point of other verbs. Studies based on eye-tracking methods, rather than language abilities, reveal signs of the ability to conceptualize mental states much earlier than 4 years old, when children can generally pass such a language-based test (Rubio-Fernández and Geurts, 2012). Bartsch and Wellman (1995) argue that the delay in the production of MSVs beyond the point of conceptualization of mental concept may result from other challenges required to refer to mental states in speech. Following such evidence, we assume here that children have the ability to conceptualize mental concepts, but still need to learn to associate MSVs with the expression of such meanings via grammatical linguistic structures.

Children also have to reach a stage of syntactic development that includes some facility
with SCs in order to fully acquire MSVs, since MSVs typically occur with SC structures to express the mental perspective on another event. At 3–5 years old, children are able to process SCs to some extent, but only imperfectly (e.g., Asplin, 2002). However, even when children are able to produce SCs with other verbs (such as verbs of communication, as in *He said Mom went home*), there is a lag before they productively use MSVs to refer to actual mental content (Diessel and Tomasello, 2001). Researchers have noted that children use MSVs in fixed phrases, in a performative use, or as a pragmatic marker, well before they use them to refer to actual mental content (e.g., Diessel and Tomasello, 2001; Shatz et al., 1983). Here by “acquisition of MSVs”, we are specifically referring to children learning usages that genuinely refer to mental content.

Another factor suggested to contribute to the difficulty of acquiring MSVs is their informational requirements (Gleitman et al., 2005; Papafragou et al., 2007). Children learn word meanings by figuring out which aspects of an observed scene are referred to by a particular word (Quine, 1960). MSVs often refer to aspects of the world that are not directly observable (i.e., the beliefs and desires of another entity). In addition to the challenges posed by their developing linguistic/conceptual abilities, children may simply have difficulty in identifying the relevant mental content necessary to learning MSVs. Moreover, this challenge is likely to be enhanced by the frequent use of MSVs to refer to more visually salient actions, such as a physical action, to express a mental perception of such events, as in “She thinks daddy is making pancakes”.

In particular, Papafragou et al. (2007) have shown that even adults who have mastered adequate conceptual and linguistic abilities attend to the physical actions in a given scene that depicts a mental event more frequently than to the mental states. Their analysis suggests that the ability to predict the usage of MSVs relies on a combination of the various cognitive and linguistic skills described above (Howard et al., 2008). Papafragou et al. have shown that the mental states are attended to at a higher rate given cues that heighten the salience of the mental content by manipulating the syntactic and semantic
properties of the given content. However, their analysis shows that the sensitivity of children to such cues lags behind that of adults. They conclude that the need to learn the associations between such cues and the relevant mental content is a factor in children’s late acquisition of MSVs.

We propose to analyze the interaction of the above factors over the trajectory of MSV acquisition via a computational model, i.e., (i) the unique semantic properties of MSVs, (ii) the use of SC to express MSVs, and (iii) the informational requirement of MSVs. While some computational models analyze the development of the cognitive ability to conceptualize mental states (Baker et al., 2011; Richardson et al., 2012), and others include SC syntax into language acquisition models (Freudenthal et al., 2002), to our knowledge, this is the first computational model that incorporates mental meanings and SC structures into a language learning model. Specifically, we extend a model of argument structure acquisition (Alishahi and Stevenson, 2008) to enable the representation and processing of MSVs. For this purpose, we enrich the input representation of the model with mental state semantic and syntactic properties, and modify the model to address the difficulty in attending to such mental concepts. We then use our model to simulate the developmental change proposed by Papafragou et al. through a gradually increasing ability to attend to the mental content of a scene, and analyze the role of this developmental change in replicating the findings of Papafragou et al. (2007).

Importantly, to account for the pattern of errors in child data, our computational learner must attend to the observed SC syntax in an MSV utterance, at the same time that the learner’s ability to attend to the mental content is limited. Our model thus extends the account of Papafragou et al. (2007) to show that a probabilistic interplay of the semantic and syntactic features of a partial and somewhat erroneous interpretation of the input, combined with a growing ability to attend to cues indicative of mental content, can help to account for children’s developmental trajectory in learning MSVs.
3.1.1 Background and Our Approach

To investigate the linguistic and contextual cues that could help in learning MSVs, Papafragou et al. (2007) use a procedure called the Human Simulation Paradigm (originally proposed by Gillette et al., 1999). In this paradigm, subjects are put in situations intended to simulate various word learning conditions of young children. For example, in one condition, adults watch silent videos of caregivers interacting with children, and are asked to predict the verb uttered by the caregiver. In another condition, subjects hear a sentence containing a nonce verb (e.g., *gorp*) in addition to displayed video, and are asked what *gorp* might mean as its used in both the sentence and referring to the video.

We focus on two factors investigated by Papafragou et al. (2007) in the performance of adults and children in identifying MSVs. The first factor they investigated involved the syntactic frame used when subjects were given a sentence with a nonce verb. Papafragou et al. (2007) hypothesized that an SC frame would be a cue to mental content (and an MSV), since the SC refers to propositional content. The second factor Papafragou et al. (2007) examined was whether the video described a “true belief” or a “false belief” scene: A true belief scene shows an ordinary situation which unfolds as the character in the scene expects — e.g., a little boy takes food to his grandmother, and she is there in the house as expected. The corresponding false belief scene has an unexpected outcome for the character — in this case, another character has replaced the grandmother in her bed. Here the hypothesis was that such false belief scenes would heighten the salience of mental activity in the scene (e.g., what the boy is thinking when he sees someone different than his grandmother). The heightened salience should, in turn, lead to greater Belief verb responses in describing the scenes.

The results of Papafragou et al. (2007) showed that both adults and children were sensitive to both the scene and syntax cues, but children’s ability to draw on such cues was inferior to that of adults in helping them to predict Belief verbs. They thus propose that the difference between children and adults is that children have not yet formed as
strong an association as adults between the cues and the mental content of a scene as required to match the performance of adults. Nonetheless, their results suggest that the participating children had the conceptual and linguistic abilities required for MSVs, since they were able to produce them under conditions with sufficiently strong cues.

In this chapter, we simulate the experiments of Papafragou et al. (2007) using a novel computational approach. Following Papafragou et al. (2007), we assume that even when a learner is able to perceive the general semantic and syntactic properties of a belief scene and associated utterance, they may not attend to the mental content in every situation, and that the ability to do so improves over time. We model a developmental change in a learner’s attention to mental content: At early stages, corresponding to the state of young children, the learner largely focuses on the action aspects of a belief scene, even in the presence of an utterance using an MSV. Over time, the learner gradually increases in the ability to attend appropriately to the mental aspects of such a scene and utterance, until adult-like competence is achieved in associating the available cues with mental content.

Importantly, our work extends the proposal of Papafragou et al. (2007) by bringing in evidence from other relevant studies on children’s ability to process SCs. More specifically, we suggest that when children hear a sentence like *I think Mom went home*, they recognize (and record) the existence of an SC, while *at the same time* they focus on the action semantics as the main (most salient) event. In other words, we assume that children’s imperfect syntactic abilities are at least sufficient to recognize the SC usage (Nelson et al., 1989; Asplin, 2002). However, their attention is mostly directed towards the action expressed in the embedded complement, either because mental content is less easily observable than actions (Papafragou et al., 2007), or due to the linguistic saliency of the embedded clause (Diessel and Tomasello, 2001; Dehe and Wichmann, 2010). As mentioned above, we model this misrepresentation by considering the possibility of not attending to mental content in a belief scene. Specifically, we assume that (i) the model is very likely to overlook the mental content at earlier stages (corresponding to children’s
observed behaviour); and (ii) as the model ‘ages’ (i.e., receives more input), its attentional abilities improve and thus the model is more likely to focus on the mental content as the main proposition. Our results suggest that these changes to the model lead to a match between our model’s behaviour and the differential results of Papafragou et al. (2007) for children and adults.

3.2 The Computational Model

Most computational models of verb argument structure acquisition have largely focused on physical action verbs (Alishahi and Stevenson, 2008; Chang, 2009; Perfors et al., 2010; Parisien and Stevenson, 2011). We present a computational model that extends the incremental Bayesian model of Alishahi and Stevenson (2008) to include the syntactic and semantic features required for the processing of MSVs and other verbs that take SCs. This model provides us with an incremental framework in which we can examine developmental patterns as it gradually learns relevant aspects of argument structures. Moreover, this model enables us to analyze the developmental changes in attention to the mental state content of an MSV usage and its semantic and syntactic representation as it learns argument structures. In this section, we first present an overview of the model (Section 3.2.1), followed by a description of the original model of Alishahi and Stevenson (Section 3.2.2), and a description of our extensions (Section 3.2.3).

3.2.1 Overview of the Model

The input to the model is a sequence of frames, where each frame is a collection of syntactic and semantic features representing what the learner might extract from an utterance s/he has heard paired with a scene s/he has perceived. In particular, we consider syntactic properties such as syntactic pattern, as well as semantic properties, including event primitives and event participants. These properties enable the specification of character-
istics that distinguish various types of usages of MSVs among other classes of verbs, e.g., expression of Desire vs. Belief meaning. Table 3.1 presents a sample frame illustrating possible values for these features.

The model incrementally groups the input frames into clusters that resemble argument structure constructions that reflect probabilistic associations of the syntactic and semantic features across similar verb usages. Each learned cluster is a probabilistic (and possibly noisy) representation of an argument structure construction: e.g., a cluster containing semantic and syntactic properties corresponding to usages such as *I eat apples*, *She took the ball*, and *He got a book*, etc., represents a Transitive Action construction.\(^1\) Such constructions allow for some degree of generalization over the observed input given similarity across the verb usages. For example, the model can predict the semantic properties of a usage of a novel verb in a transitive utterance such as *She gorped the ball*, by estimating the most likely semantic properties for observed verbs in a similar syntactic and semantic context. Such probabilistic reasoning is especially powerful because clusters involve complex interactions of features, and the model reasons across all such clusters to make suitable generalizations over its learned knowledge.
Chapter 3. Attending to Mental State Verbs

3.2.2 Learning Constructions of Verb Usages

This section describes the model of Alishahi and Stevenson (2008) that we use as a basis for our model. The model of Alishahi and Stevenson groups input frames into clusters on the basis of the overall similarity in the values of their syntactic and semantic features. Importantly, the model learns these clusters incrementally; the number and type of clusters is not predetermined. The model considers the creation of a new cluster for a given frame if the frame is not sufficiently similar to any of the existing clusters. For an illustration of the clusters created by the model please refer to Figure 3.1. Formally, the model finds the best cluster for a given input frame \( F \) as in:

\[
\text{BestCluster}(F) = \arg \max_{k \in \text{Clusters}} P(k|F)
\]  

(3.1)

Note that, because the associations are probabilistic, an argument structure construction may be represented by more than one cluster in our model.
where $k$ ranges over all existing clusters and a new one. Using Bayes rule:

$$P(k|F) = \frac{P(k)P(F|k)}{P(F)} \propto P(k)P(F|k) \quad (3.2)$$

$P(F)$ is constant for all clusters in respect to the given frame and therefore is dropped in 3.2. The prior probability of a cluster $P(k)$ is estimated as the proportion of frames $n_k$ that are in $k$ out of all observed input frames, $n$ thus assigning a higher prior to larger clusters, representing more frequent constructions.\(^2\) Formally:

$$P(k) = \frac{n_k}{n + 1} \quad (3.3)$$

We take the normalizing factor to be $n + 1$ to account for the possibility of creating a new cluster, containing the new frame, with the probability:

$$P(k_0) = \frac{1}{n + 1} \quad (3.4)$$

The likelihood $P(F|k)$ in Equation (3.2) is estimated based on the match of feature values in $F$ and in the frames of $k$ (assuming independence of the features):

$$P(F|k) = \prod_{i \in \text{Features}} P_i(j|k) \quad (3.5)$$

where $i$ refers to the $i^{th}$ feature of $F$ and $j$ refers to its value, and $P_i(j|k)$ is calculated as:

$$P_i(j|k) = \frac{\text{count}_i(j,k) + \lambda}{n_k + \lambda \alpha_i} \quad (3.6)$$

\(^2\)Equation (3.3) is based on the formula used by Anderson (1991), $P(k) = \frac{c n_k}{(1-c) n + c n}$, and similarly for Equation (3.4), where $c$ is a coupling factor used to determine the likelihood of grouping a new item into an existing cluster. Following Alishahi and Stevenson (2008), we set $c = 0.5$ to minimize the degree of bias of our model in its clustering decisions.
where \( \text{count}_i(j,k) \) is the number of times feature \( i \) has the value \( j \) in cluster \( k \), and \( n_k \) is the number of frames in \( k \). The parameter \( \lambda \) is set to a small value (\( \lambda \ll 1 \)) to assign a non-zero probability to clusters that have not yet recorded an occurrence of feature \( i \) with the value \( j \). Notably, \( \lambda \) becomes less significant to the overall value of \( P_i(j|k) \) as a cluster increases in size relative to the normalizing factor here, which is dependent on the size of the cluster, \( n_k \). Therefore, the parameter \( \lambda \) is more effective for newer clusters that still record low values of \( \text{count}_i(j,k) \). The parameter \( \lambda \) compensates for the yet to acquire knowledge of such newer clusters that are likely to observe more values for each feature as they would accumulate knowledge over additional input frames. The smoothing parameter \( \alpha_i \) estimates the number of possible values for each feature \( i \). The \( \alpha_i \) parameter would provide a non-zero probability when calculating the probability of a new cluster, by giving equal likelihood for each feature value as \( \frac{1}{\alpha_i} \).

The above calculation for the count\(_i\) operator as the number of occurrences of the value \( j \) in cluster \( k \) holds for most of the features in the data that can be described as a single-value feature (e.g., predicate, syntactic pattern, etc.). However, the data also includes set-valued features such as, event participants, which consist of a set of values per each occurrence of the feature. For such features, exact comparison is too strict. The preferred comparison method should consider the degree of overlap between the sets of values to allow a more flexible similarity measure. For example, such a method would take into account that think and see both refer to a state rather than an action, though one refers to a mental event whereas the other does not. We modify Equation (3.6) for the set-valued features as follows:

\[
P^\text{set}_i(j|k) = \frac{\sum_{f \in k} \text{sem-score}(S_j, S_f) + \lambda |S_j|}{\sum_{l \in \text{lexicon}} \sum_{f \in k} \text{sem-score}(S_l, S_f) + \lambda \alpha_i^\text{set}}
\]

where \( j \) is the value for feature \( i \) represented by the set \( S_j \). \( \text{sem-score}(S_j, S_f) \) measures the degree of similarity between two sets of features: \( S_j \) and \( S_f \). \( S_f \) ranges over any other
possible set of values for the feature \( i \) for any of the frames \( f \) included in cluster \( k \). Parallel to Equation (3.6), we substitute count\(_i\) with a sum over the similarity scores between the set \( S_j \) and any other set of values represented in cluster \( k \). Here, the smoothing parameter \( \alpha^\text{set}_i \) estimates the number of sets of values this feature may hold. We measure the degree of similarity using the Jaccard similarity score:

\[
\text{sem-score}(S_j, S_f) = \frac{|S_j \cap S_f|}{|S_j \cup S_f|}
\]

(3.8)

By using this similarity measure, the degree of similarity of two sets is increased by each shared feature, while simultaneously penalized for unshared values through the denominator. Equation (3.7) thus estimates the probability of a set \( S_j \) to correspond to the observed set values in a given cluster \( k \) by summing over its partial similarity to the sets observed in the cluster over all the possible sets of values for this feature in the denominator of Equation (3.7).

### 3.2.3 Modeling Developing Attention to Mental Content

One factor proposed to play an important role in the acquisition of MSVs is the difficulty children have in being aware of (or perceiving the salience of) the mental content of a scene that an utterance may be describing (Papafragou et al., 2007). This difficulty arises because the aspects of a scene associated with an MSV — the “believing” or the “wanting” — are not directly observable, as they involve the inner states of an event participant. Instead, younger children tend to focus on the physical (observable) parts of the scene, which generally correspond to the event described in the embedded clause of an MSV utterance. For instance, young children may focus on the “going” action in *I think mom went home*, rather than on the “thinking”.

A key component of our model is a mechanism that simulates the gradually-developing ability in children to attend to the mental content rather than solely to the (embedded)
### Scene–Utterance Input Pair:

\[
\begin{align*}
\text{Think}\{\text{state, consider, cogitate}\} & [I\{\text{experiencer, preceiver, considerer}\}, \\
\text{Go}\{\text{physical, act, move}\} & (\text{MOM}\{\text{agent, change}\}, \text{HOME}\{\text{location, destination}\})
\end{align*}
\]

*I think Mom went home.*

#### Extracted Frames:

- **Interpretation#1 (mental event is attended to)**
  - **main predicate**: think
  - **other predicate**: go
  - **event primitives**: \{state, consider, cogitate\}
  - **event participants**: \{experiencer, preceiver, considerer\}
    - \{preposition, action, perceivable\}
  - **syntactic pattern**: arg1 verb arg-S
  - **verb count**: 2

- **Interpretation#2 (mental event not attended to)**
  - **main predicate**: go
  - **other predicate**: think
  - **event primitives**: \{physical, act, move\}
  - **event participants**: \{agent, change\}
    - \{location, destination\}
  - **syntactic pattern**: arg1 verb arg-S
  - **verb count**: 2

Table 3.2: A scene–utterance pair and the two frames extracted from it. The bottom two panels of the table describe the two possible interpretations given the input pair. (a) Interpretation#1 assumes that the mental event is the focus of attention. Here think is interpreted as the main predicate, which the event primitives and participants refer to. (b) Interpretation#2 assumes that attention is mostly directed to the physical action in the scene, and thus go is taken to be the main predicate, which also determines the extracted event primitives and participants. Note that for both interpretations, the learner is assumed to perceive the utterance in full, thus both verbs are heard in the context of the sentential complement syntax (i.e., syntactic pattern with SC and 2 verbs), without fully extracting the syntactic relations between the clauses.
physical action. This mechanism basically entails that the model may “misinterpret” an input frame containing an MSV as focusing on the semantics of the action in the sentential complement. To adapt the model in these ways, we change the frame extraction component to allow two possible interpretations for a mental event input. First, to reflect the proposal of Papafragou et al. (2007), we incorporate a mechanism into the model’s frame-extraction process that takes into account the probability of attending to mental content. Specifically, we assume that when presented with an input pair containing an MSV, as in Table 3.2, a learner attends to the perceptually salient action/state expressed in the complement (here go) with probability $p$, and to the non-perceptually salient mental event expressed in the main verb (here think) with probability $1 - p$. This probability $p$ is a function over time, corresponding to the observed developmental progression. At very early stages, $p$ will be high (close to 1), simulating the much greater saliency of physical actions compared to mental events for younger children. With subsequent input, $p$ will decrease, giving more and more attention to the mental content of a scene with a mental event, gradually approaching adult-like abilities.

We adopt the following function for $p$:

$$p = \frac{1}{\delta \cdot t + 1}, \quad 0 < \delta \ll 1$$

(3.9)

where $t$ is the current time, expressed as the total number of scene–utterance pairs observed thus far by the model, and the parameter $\delta$ is set to a small value to assign a high probability to the physical action interpretation of the scene in the initial stages of learning (when $t$ is small).

We must specify the precise make-up of the frames that correspond to the two possible interpretations considered with probability $p$ and $1 - p$. Papafragou et al. (2007) state only that children and adults differentially attend to the action vs. mental content of the scene. We operationalize this by forming two possible frames in response to an MSV
usage. We propose that one of the frames (with probability $1 - p$) is the interpretation of the mental content usage, as in Table 3.2(a). However, we extend the account of Papafragou et al. (2007) by proposing that the other frame considered is not simply a standard representation of an action scene–utterance pair. Rather, we suggest that the interpretation of an MSV scene–utterance pair that focuses on the action semantics does so within the context of the SC syntax, given the assumed stage of linguistic abilities of the learner. This leads to the frame (with probability $p$) as in Table 3.2(b), which represents the action semantics within a two-verb construction associated with the SC syntax.

3.3 Experimental Setup

3.3.1 Input Data

We generate artificial corpora for our simulations, since we do not have access to sufficient data of actual utterances paired with scene representations. In order to create naturalistic data that resembles what children are exposed to, we follow the approach of Alishahi and Stevenson (2008) to build an input-generation lexicon that has the distributional properties of actual CDS. Their original lexicon contains only high-frequency physical action verbs that appear in limited syntactic patterns. Our expanded lexicon also includes mental state, Perception, and Communication verbs, all of which can appear with SCs.

We extracted our verbs and their distributional properties from the child-directed speech of 8 children in the CHILDES database (MacWhinney, 2000). We selected 28 verbs from different semantic classes and different frequency ranges: 12 physical action verbs taken from the original model (come, go, fall, eat, play, get, give, take, make, look, put, sit), 6 Perception and Communication verbs (see, hear, watch, say, tell, ask), 5 Belief

---

3 Corpora of Brown (1973); Suppes (1974); Kuczaj (1977); Bloom et al. (1974); Sachs (1983); Lieven et al. (2009).
Table 3.3: A sample of the lexical entries for the verb *know* in our input generation lexicon. Each verb entry consists of its overall frequency and a list of frames. The entry specifies the overall frequency of the verb in the data according to all its usages in the 8 CHILDES corpora included. This frequency is used to randomly draw the verb for each usage included in the input to mimic the distribution of verb in CDS. The semantic and syntactic properties of the verb usage are then drawn from the list of frames based on the frequency of the frame in the annotated sample.
verbs (think, know, guess, bet, believe), and 5 Desire verbs (want, wish, like, mind, need). For each verb, we manually analyzed a random sample of 100 CDS usages (or all usages if fewer than 100) to extract distributional information about its argument structures.

We construct the input-generation lexicon by listing each of the 28 verbs (i.e. the ‘main predicate’), along with its overall frequency, as well as the frequency with which it appears with each argument structure. Each entry contains values of the syntactic and semantic features (see Table 3.3 for examples), including ‘event primitives’, ‘event participants’, ‘syntactic pattern’, and ‘verb count’.

By including these features, we assume that a learner is capable of understanding basic syntactic properties of an utterance, including word syntactic categories (e.g., noun and verb), word order, and the appearance of SCs (e.g., Nelson et al., 1989). We also assume that a learner has the ability to perceive and conceptualize the general semantic properties of events — including mental, perceptual, communicative, and physical actions — as well as those of the event participants. Values for the semantic features (the event primitives and event participants) are taken from Alishahi and Stevenson (2008) for the action verbs, and from several sources including VerbNet (Kipper et al., 2008) and Dowty (1991) for the additional verbs.

For each simulation in our experiments (explained below), we use the input-generation lexicon to automatically generate an input corpus of scene–utterance pairs that reflects the observed frequency distribution in CDS.

3.3.2 Setup of Simulations

Our goal here is to evaluate the ability of our model to predict the use of an MSV given various input scenarios following the experimental settings of Papafragou et al.\footnote{Note that we only include frames that occurred more than once in the sample to focus on the frequent usage types for each verb.} The model does not use the input-generation lexicon in learning.\footnote{The model does not use the input-generation lexicon in learning.}
### Table 3.4: The three test scenarios simulating the experimental settings used by Papafragou et al. (2007). The table specifies the experimental condition that corresponds to the simulated condition, together with the semantic and/or the syntactic features used to simulate the condition.

<table>
<thead>
<tr>
<th>Simulated Scenario</th>
<th>Psycholinguistic Scenario</th>
<th>Frame Make-up</th>
<th>Syntactic Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene-only</td>
<td>Subjects watch a silent video depicting either an Action or a Belief scene.</td>
<td>Event primitives and event participants corresponding to the scene type</td>
<td>None</td>
</tr>
<tr>
<td>Syntax-only</td>
<td>Subjects hear either an SC or a non-SC utterance.</td>
<td>None</td>
<td>Syntactic pattern and verb count of the utterance type heard</td>
</tr>
<tr>
<td>Syntax &amp; scene</td>
<td>Subjects watch a silent video (with Action or Belief content), and hear an associated (non-SC or SC) utterance.</td>
<td>Event primitives and event participants corresponding to the scene type</td>
<td>Syntactic pattern and verb count of the utterance type heard</td>
</tr>
</tbody>
</table>

As mentioned above, Papafragou et al. (2007) present several sets of experiments using the Human Simulation Paradigm presented in Section 3.1.1. Their experiments include: (i) verb prediction task with adult participants, and (ii) verb prediction task with child and adult participants. In the first experiment, adult participants were presented with input that corresponds to each of the three scenarios presented in Table 3.4. The scenarios were coupled with two input conditions: typical Action context (i.e., physical action scene and transitive syntax), or Belief context (i.e., a belief scene and finite-SC syntax). The participants were evaluated on their ability to predict Belief verbs given each context scenario and condition. Papafragou et al. (2007) compare the prediction rate of Belief verbs across input types assuming higher prediction rates of Belief verb would be observed given the Belief input. In order to evaluate the prediction rate of Belief verbs, they measure the percentage of Belief verbs responses out of all predictions.

In the second experiment, Papafragou et al. (2007) focus on the syntax & scene
scenario, but compare the prediction rate of children and adults given the two input conditions, Action and Belief input. In addition, the results here include the prediction rate for several verb classes such as, Action and Belief. The prediction rate for each verb class is calculated by collapsing all the verb predictions that correspond to each of the studied verb classes into one group. Again, Papafragou et al. (2007) predict that higher prediction rate of Belief verbs would be observed given the Belief content. We first describe our simulation of their experiments, followed by the results of their experiments and our simulations Section 3.4.

We perform simulations of the verb prediction task in the human simulation paradigm as follows: At each test point, we present the model with a partial test frame with missing predicate (verb) values, and different amounts of information for the other features. The tests correspond to the scenarios in the original experiments of Papafragou et al. (2007), where each scenario is represented by a partial frame as described in Table 3.4. We perform simulations by training the model on a randomly generated input corpus, and examining changes in its performance over time with periodic tests. The periodical test points enable the simulation of the prediction abilities of child and adult participants.

We perform 100 simulations, each on 15000 randomly-generated training frames, and examine the type of verbs that the model predicts in response to test frames for the three scenarios. For each scenario and each simulation, we generate a test frame by including the relevant feature values of a randomly-selected physical action or belief verb usage from the input-generation lexicon.

Papafragou et al. (2007) code the individual verb responses of their human subjects into various verb classes (e.g., Action and Belief). To analogously code our model’s response to each (partial) test frame, we estimate the likelihood of each of two verb groups, Belief and Action, by summing over the likelihood of all the verbs in that group. In the results below, these likelihood scores are averaged for each test point over the 100

---

6The Action verbs include Action, Communication, and Perception verbs, as by Papafragou et al. (2007). Verbs from the Desire group are not considered here, also as in (Papafragou et al., 2007).
Given a test frame $F_{\text{test}}$, we use the clusters learned by the model to calculate the likelihood of each of the 31 verbs $v$ as the response of the model indicating the meaning of the novel verb, as in:

$$P(v|F_{\text{test}}) = \sum_{k \in \text{Clusters}} P_{\text{head}}(v|k)P(k|F_{\text{test}})$$

$$\propto \sum_{k \in \text{Clusters}} P_{\text{head}}(v|k)P(F_{\text{test}}|k)P(k)$$

where $P_{\text{head}}(v|k)$ is the probability of the head feature having the value $v$ in cluster $k$, calculated as in Equation (3.6); $P(F_{\text{test}}|k)$ is the probability of the test frame $F_{\text{test}}$ given cluster $k$, calculated as in Equation (3.5); and $P(k)$ is the prior probability of cluster $k$, calculated as explained in Section 3.2.

What we really want to know is the likelihood of the model producing a verb from a given semantic class, rather than the likelihood of any particular verb. For each test frame, we calculate the likelihood of a semantic class by summing the likelihoods of the verbs in that class:

$$P(\text{Class}|F_{\text{test}}) = \sum_{v_c \in \text{Class}} P(v_c|F_{\text{test}})$$

where $v_c$ is one of the verbs in Class, and Class can be either the Action or the Belief class.

When our model is presented with a test frame containing a Belief scene, we assume that the model (like a language learner) may or may not attend to the mental content, resulting in one of the two interpretations described in Section 3.2.3 (see Table 3.2). To simulate this possibility, we present to the model two frames corresponding to each of the two interpretations to simulate the perception of the Belief scene. We thus calculate the verb class likelihoods using a weighted average of the verbs predicted under the two simulations.
interpretations given each of the frames. That is, we calculate the likelihood of each verb class as in Equation (3.10) for each of the two test frames.

\[
P(\text{Belief}|F_{\text{test}}) = p_i \cdot P(\text{Belief}|F_{\text{mental}}) + (1 - p_i) \cdot P(\text{Belief}|F_{\text{non-mental}})
\]

where \(F_{\text{mental}}\) corresponds to Interpretation#1 in Table 3.2 (i.e., in which the mental event is attended to), and \(F_{\text{non-mental}}\) corresponds to Interpretation#2 in Table 3.2 (i.e., in which the mental event is not attended to). \(p_i\) is the probability that the mental interpretation is perceived given to the Belief scene.

Following Papafragou et al. (2007), we test our model with two types of Belief scenes: True Belief and False Belief, with the latter having a higher level of belief saliency. We model the difference between these two scene types as a difference in the probabilities of perceiving the two interpretations, with a higher probability for the belief interpretation given a False Belief test frame. In the experiments presented here, we set this probability to 80% for False Belief, and to 60% (just above chance) for True Belief. (Unlike in training, where we assume a change over time in the probability of a belief interpretation, for each presentation of the test frame we use the same probabilities of the two interpretations.)

### 3.4 Experimental Results

We present two sets of results: In Section 3.4.1, we examine the role of syntactic and semantic cues in MSV identification, by comparing the likelihoods of the model’s Belief verb predictions across the three scenarios (see Table 3.4). We test the model after processing all 15000 input frames, assuming this simulates an adult-like behaviour (as observed in adult participants by Papafragou et al. (2007)). At this stage, we present the model with an Action test frame (Action scene and/or Transitive syntax), or a Belief
test frame (False Belief scene and/or SC syntax simulating the heighten Belief content presented by Papafragou et al. (2007)). In Section 3.4.2, we look into the role of semantic cues that enhance belief saliency, by comparing the likelihoods of Belief vs. Action verb predictions in the syntax & scene scenario. The test frames depict either a True Belief or a False Belief scene, paired with an SC utterance. Here, we test our model periodically to examine the developmental pattern of MSV identification, comparing our results with the difference in the behaviour of children and adults observed by Papafragou et al. (2007).

3.4.1 Linguistic Cues for Belief Verb Prediction

Following prior research showing a low rate of Belief verbs in Human Simulation Paradigm studies, Papafragou et al. (2007) initially evaluated the prediction rate of Belief verbs in response to enhanced Belief content in comparison to Action content as described in Section 3.3.2. The results of Papafragou et al. (2007), presented in Figure 3.2, left side, show a significant influence of content type. The highest percentage of Belief verbs predicted was based on the syntax & scene scenario, while the syntax-only scenario promoted higher rate of Belief verbs than the semantics-only scenario. This result emphasizes the

Figure 3.2: Likelihood of Belief verb prediction given Action or Belief input.
importance of syntactic information for MSV prediction.

The right side of Figure 3.2 shows the likelihood of Belief verb prediction by our model. Similar to the results of Papafragou et al. (2007), our model’s likelihood of Belief verb prediction is extremely low when given an Action test frame (Action scene and/or Transitive syntax), whereas it is much higher when the model is presented with a Belief test frame (False Belief scene and/or SC syntax). Moreover, as in Papafragou et al. (2007), when the model is tested with Belief content, the lowest likelihood is for the scene-only scenario and the highest is for the syntax & scene scenario.

Papafragou et al. (2007) found, somewhat surprisingly, that the syntax-only scenario was more informative for MSV prediction than the scene-only scenario. Our results replicate this finding. Given the unique semantic properties of Belief verbs, the scene-only scenario would be expected to be more informative. However, the associations formed by our model over the constructions provide an explanation for this phenomenon. Non-SC usages of MSVs, such as, *I believe you*, are often grouped with Action verbs that frequently appear with non-SC syntax, and this results in constructions with mixed (Action and Belief) semantics. When using MSV semantic features to make the verb prediction, the Belief verbs get a lower likelihood compared with the other input scenarios based on such mixed constructions. However, the frequent usage of MSVs with SC results in entrenched constructions of mostly MSVs. Although other verbs, such as *see* and *say*, may also be used with SC syntax, they are grouped with verbs such as *watch* and *tell* into constructions with mixed (SC and non-SC) syntax. When given SC syntax in verb prediction, the more coherent MSV constructions result in a high likelihood of predicting Belief verbs.

These results imply that children may be able to predict the use of MSVs give an SC utterance with some confidence due to the stronger entrenchment of the MSV-SC construction compared with the constructions of non-MSVs and SC. At the same time, children may present inferior ability to predict the use of MSVs from a given semantic
Figure 3.3: Verb class likelihood: (a) The results of Papafragou et al. (2007) for adults and children (aged 3;7–5;9); (b) Model’s results given True Belief; (c) Model’s results given False Belief. Action verb class corresponds to Action, Communication and Perception verbs as in (Papafragou et al., 2007). Verbs from the Desire verb class are not presented here.

Scene–Utterance Input Pair:

Think\_[state, consider, cogitate]_[(I\_[experiencer, preceiver, considerer], Go\_[physical, act, move]_[(MOM\_[agent, change],

HOME\_[location, destination)])]

I think Mom went home.

Extracted Physical Frame:

<table>
<thead>
<tr>
<th>main predicate</th>
<th>go</th>
</tr>
</thead>
<tbody>
<tr>
<td>other predicate</td>
<td>none</td>
</tr>
<tr>
<td>event primitives</td>
<td>{physical, act, move}</td>
</tr>
<tr>
<td>event participants</td>
<td>{agent, change}</td>
</tr>
<tr>
<td></td>
<td>{location, destination}</td>
</tr>
<tr>
<td>syntactic pattern</td>
<td>arg1 verb arg2</td>
</tr>
<tr>
<td>verb count</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.5: The frame representing a physical interpretation of the scene–utterance pair presented also in Table 3.2. The interpretation assumes that the physical event is the focus of attention. Here go is interpreted as the main predicate, which the event primitives and participants refer to. Here, the learner is assumed to perceive only the complement clause of the utterance.

content as explained by the stronger association of non-MSVs, such as Action verbs, to non-SC syntax. This hypothesis may be validate via additional psycholinguistic experimental work in the future, since Papafragou et al. (2007) do not include prediction results from children given the three input scenarios.

3.4.2 Belief Saliency in Verb Prediction

Following their result from the above experiment, Papafragou et al. (2007) compare the verb predictions provided by children aged 3;7 to 5;9 and adults when presented only
Chapter 3. Attending to Mental State Verbs

with the syntax & scene scenario (see Section 3.3.2). Papafragou et al. (2007) also focus here on conditions where the scene included belief content, i.e. scene data was drawn from True Belief (TB) and False Belief (FB) scenes and accompanied by an utterance with an SC.

Figure 3.3(a) shows the results of Papafragou et al. (2007) for children and adults, for True Belief and False Belief scene input. Figures 3.3(b) and (c) present the likelihoods of the model’s Belief vs. Action verb prediction, over time, for True and False Belief situations (with SC syntax and True/False Belief scene), respectively. We first compare the responses of our model at the final stage of training to those of adults in Papafragou et al. (2007). At this stage, the model’s verb predictions (for both True and False Belief) follow a similar trend to that of adult subjects in Papafragou et al. (2007). The likelihood of Belief verbs is much higher than the likelihood of Action verbs given a False Belief situation. Moreover, the likelihood of Belief verbs is higher given a False Belief situation, compared to a True Belief situation.

Next, we compare the developmental pattern of Belief/Action verb predictions in the model with the difference in behaviour of children and adults in Papafragou et al. (2007). We focus on the model’s responses after processing about 3000 input pairs, as it corresponds to the trends observed for the children in Papafragou et al. (2007). At this stage, the likelihood of Belief verbs is lower than that of Action verbs for the True Belief situation, but the pattern is reversed for False Belief, a pattern similar to children’s behaviour in Papafragou et al. (2007) (see Figure 2(a)). As in Papafragou et al. (2007), the likelihood of Belief verb predictions in our model is higher than that of Action verbs for the False Belief situation, in both “child” and “adult” stages, with a larger difference as the model ‘ages’ (i.e., processes more input). For the True Belief situation also the pattern is similar to that of Papafragou et al. (2007): Belief verbs are less likely than Action verbs to be predicted at early stages, but as the model receives more input, the likelihood of Belief verbs becomes slightly higher than that of Action verbs.
Papafragou et al. (2007) hypothesize that greater attention to the Action content of a scene indirectly leads children to focus on the Action semantics and syntax of the embedded SC of a Belief verb. We have suggested instead that the focus is on the Action semantics within the context of the SC syntax of the MSV. To directly evaluate the necessity of our latter assumption (that the learner notes the SC usage), we perform an additional simulation that represents a situation in which children fully interpret the scene as representing a physical action while disregarding the use of SC. The non-SC physical interpretation frame is essentially equivalent to a frame representing a usage of an Action verb, in this case *go*, since it does not record the additional MSV being used, here *think*, nor does it records the use of an SC. We aim to confirm that disregarding the use of an SC would not lead to a delayed acquisition pattern observed in the simulation above. For this purpose, we use both Action syntax and semantics to represent the non-SC physical interpretation of the Belief scene in both the training and testing (see Table 3.5). Based on these settings, the model predicts high likelihood for the Belief verbs from a very early stage, not showing the same delayed acquisition pattern exhibited by the results of Papafragou et al. (2007). In early training stages, the likelihood of attending to the mental content in usages of SCs is low. Yet, the few usages of MSVs with SCs that are correctly recorded lead to a strong association between MSVs and the SC syntax. The high frequency of MSVs with the SC syntax compared with other verbs, such as *see* and *say* results in entranced constructions of MSVs with SC usage, even though they are based on a relatively low number of such usages. This result suggests that the SC syntax plays an important role in MSV acquisition even when children are not attending to the mental content.
3.5 Discussion

Various studies have considered why mental state verbs (MSVs) appear relatively late in children’s productions (e.g., Shatz et al., 1983; Bartsch and Wellman, 1995). Several cognitive and linguistic skills have been argued to play a role in this delay, including the need to attend to the hard-to-observe mental content in order to learn its association to MSVs (Papafragou et al., 2007; Howard et al., 2008). The Human Simulation Paradigm has revealed that even adult participants have difficulties in attending to mental content since they tend to focus on the physical action cues of a scene (Gleitman et al., 2005). The results of Papafragou et al. (2007) further show that cues emphasizing mental content lead to a significant increase in MSV responses in such tasks from both children and adults. Moreover, they show that (for adults) a sentential complement (SC) structure is a stronger cue to an MSV than the semantic cues emphasizing mental content.

In this chapter we adapt a computational Bayesian model to study how a learner might identify such semantic and syntactic cues. We incorporate an attentional mechanism into the model that simulates the growing sensitivity to mental content in a scene. We show that both the ability to observe mental content and the ability to recognize the use of an SC structure are essential to replicate the observations of Papafragou et al. (2007). Moreover, our results predict the strong association of MSVs to the SC syntax, for the first time (to our knowledge) in a computational model.

Children often use verbs other than MSVs in experimental settings in which MSVs would be the appropriate or correct verb choice (Asplin, 2002; Kidd et al., 2006; Papafragou et al., 2007). Our model presents similar verb choice across Belief and Action verbs. One underlying cause of this behaviour in the model is its association of action semantics to SC syntax, due to the tendency to focus on the physical cues in a scene associated with the use of an MSV with an SC. We hypothesize that the frequent use of non-MSVs, such as Perception and Communication verbs, with the SC syntax may also contribute to the initial higher prediction rate of Action verbs (see de Villiers, 2005, for
theoretical support). We further analyze this hypothesis in the following chapters.
Chapter 4

The Facilitative Role of Distributional Properties of MSVs

4.1 Introduction

Usage-based theories of language acquisition suggest that children learn argument structure regularities mainly from the input they receive. These theories are supported by observing that children initially learn verb argument structures on an item-by-item basis, and only later generalize their verb-specific knowledge into abstract constructions that map a particular syntactic form to certain semantic properties (Tomasello, 2000). The distributional properties of verb usages in child-directed input highly affect the developmental path of the acquisition of argument structure constructions. For example, several studies have shown that children tend to learn high-frequency verbs earlier (Naigles and Hoff-Ginsberg, 1998; Matthews et al., 2005), and that they are more likely to detect grammatical anomalies in sentences containing such verbs (Theakston, 2004; Ambridge et al., 2008). Moreover, the relative frequency of a verb with a particular syntactic construction has been shown to positively correlate with the ability of young children to recall sentences containing that verb in that construction (Kidd et al., 2006, 2010). Most
importantly, there is evidence that the acquisition of a construction is connected to a high-frequency exemplar verb; e.g., *give* is the exemplar for the ditransitive construction (Goldberg, 1999; Kidd et al., 2006, 2010). In fact, several studies have shown that it is not just the amount of overall exposure to a construction that affects its acquisition, but instead learning seems to be facilitated by a high number of usages of a particular exemplar verb (Casenhiser and Goldberg, 2005; Wonnacott et al., 2008).

In this chapter, we analyze how such distributional properties of the input can facilitate or hinder the acquisition of verbs, and in particular the acquisition of MSVs (Barak et al., 2013a,b). We do so while also considering two sub-groups of MSVs, i.e., Belief and Desire verbs, to allow more detailed analysis of patterns in the acquisition of MSVs as a group, and of Belief vs. Desire verbs in particular. The distinction between Belief and Desire verbs is especially essential given the observed earlier acquisition of Desire verbs compared with Belief verbs (Bartsch and Wellman, 1995; Asplin, 2002; Perner et al., 2003; Pascual et al., 2008). Importantly, Belief and Desire verbs differ in their distribution of syntactic properties, as well as fine-grained semantic properties. Desire verbs frequently occur with the infinitival-SC, e.g., “Chris wants to eat ice-cream”, and can refer to future unfulfilled events (e.g., Chris has not eaten any ice-cream yet). Belief verbs frequently occur with the finite-SC, e.g., “I think Chris is eating ice-cream”, and can be used in contrast to reality (e.g., Chris could be eating a cake while the above example is uttered). We aim to incorporate these fine-grained properties of Belief and Desire verbs into our analysis of the possible role such properties play in the acquisition process. To facilitate the analysis of the difference in the syntactic requirements of Belief and Desire verbs, we modify the data representation to distinguish between the two types of SCs and the properties of their usage, e.g., by adding a complementizer feature that differs in possible values of finite-SC and infinitival-SC (e.g., *that*, *to*, etc.). Moreover, we include in our input other verb classes that frequently occur with SCs (e.g., Communication and Perception verbs) to analyze the emergence of the strong association between MSVs and
SCs given the occurrence of other verbs with similar syntactic patterns.

The remainder of this chapter is organized as follows: we first present the modifications to the input representation that allows us to study the acquisition of MSVs given fine-grained properties of their distribution over syntactic and semantic properties (see Section 4.2). We carry out two sets of experiments using the modified representation. First, we study the emergence of an exemplar verb for the finite-SC constructions with Belief verbs while making the distinction between verbs that occur with finite-SC and verbs that occur with other SC types, e.g., infinitival-SC (see Section 4.3). Secondly, we analyze the acquisition of Belief vs. Desire verbs given their distinctive semantic and syntactic properties (see Section 4.4).

4.2 Input Representation

The computational model presented in Chapter 3 addresses some developmental patterns of MSVs overall, but it does not account for the difference between finite and infinitival SCs nor between Desire and Belief verbs. In this chapter, we use the same model as in Chapter 3, which we adapt for an additional set of experiments. However, we modify the representation of the input to enable our investigation of the Belief verbs specifically, as well as the differences among the MSV classes.

The values for syntactic features of the input are based on simple observation of the order and number of verbs and arguments in the usage, and, if an argument is an SC, whether it is finite or infinitival. We add this latter feature (the type of the SC) to the syntactic representation used in Chapter 3 to (i) enable the analysis of the acquisition of the finite-SC construction, and (ii) allow distinction of the syntactic properties associated with Desire and Belief verbs. Note that this feature does not incorporate any potential level of difficulty in processing an infinitival vs. finite SC; the feature simply records that there are three different types of embedded clausal arguments: SC-inf, SC-fin, or none.
Belief frame:

<table>
<thead>
<tr>
<th>head predicate</th>
<th>think</th>
</tr>
</thead>
<tbody>
<tr>
<td>other predicate</td>
<td>make</td>
</tr>
</tbody>
</table>

**Syntactic Features:**
- syntactic pattern: arg1 verb arg2 verb arg3
- argument count: 3
- complement type: SC-fin

**Semantic Features:**
- event primitives: \{state, consider, cogitate, action\}
- event participants: \{experiencer, perceiver, considerer\}, \{agent, animate\}, \{theme, changed\}

Desire frame:

<table>
<thead>
<tr>
<th>head predicate</th>
<th>want</th>
</tr>
</thead>
<tbody>
<tr>
<td>other predicate</td>
<td>go</td>
</tr>
</tbody>
</table>

**Syntactic Features:**
- syntactic pattern: arg1 verb to verb arg2
- argument count: 2
- complement type: SC-inf

**Semantic Features:**
- event primitives: \{state, desire, cogitate, action\}
- event participants: \{experiencer, desirer, want − state\}, \{location, destination\}

Table 4.1: An example input frame. The Syntactic features of the Belief frame reflect an utterance such as *He thinks Mom made pancakes*: i.e., syntactic pattern ‘arg1 verb arg2 verb arg3’, 3 arguments, and finite SC. The Semantic features reflect a corresponding conceptualized belief event with a physical action described in the SC (\{state, consider, cogitate, action\}) whose ‘arg1’ participant (\{experiencer, perceiver, considerer\}) perceives the ‘arg2’ (\{agent, animate\}) acting on the ‘arg3’ (\{theme, changed\}). Similarly, the features of the Desire frame reflect an utterance such as *I want to go home*
Thus, while Desire and Belief verbs that typically occur with an SC-inf or SC-fin have a distinguishing feature, there is nothing in this representation that makes Desire verbs inherently easier to process. This syntactic representation reflects our assumptions that a learner: (i) understands basic syntactic properties of an utterance, such as syntactic categories (e.g., noun and verb) and word order; and (ii) distinguishes between a finite complement, as in *He thinks that mom left*, and an infinitival, as in *He wants mom to leave*.

The values for the semantic features of a verb and its arguments are based on a simple taxonomy of event and participant role properties adapted from several resources, including Alishahi and Stevenson (2008), Kipper et al. (2008), and Dowty (1991). In particular, we assume that the learner is able to perceive and conceptualize the general semantic properties of different kinds of events (e.g., *state* and *action*), as well as those of the event participants (e.g., *agent*, *experiencer*, and *theme*). In an adaptation of the lexicon used in Chapter 3, we make minimal assumptions about shared semantics across verb classes following the distinctions suggests by Dowty (1991) and in VerbNet Kipper et al. (2008). Specifically, to encode suitable semantic distinctions among MSVs, and between MSVs and other verbs, we aimed for a representation that would capture reasonable assumptions about high-level similarities and differences among the verb classes. As with the syntactic features, we ensured that we did not simply encode the result we are investigating (that children have facility with Desire verbs before Belief verbs) by making the representation for Desire verbs easier to learn. Table 4.1 presents a sample frame illustrating possible values for the syntactic and semantic features. Note that Belief and Desire verb differ in their semantic properties, i.e., having a *consider* or *desire* primitive, as shown in Table 4.1. However, as mentioned above, these primitives do not simply encode the two mental groups, Belief and Desire, since, for example, the various Belief verbs may also differ in their specific make-up of *event primitives* for each verb by possibly including primitives such as, *believer* and *estimator*. 

Chapter 4. The Facilitative Role of Distributional Properties of MSVs
4.3 Modeling the Emergence of an Exemplar Verb in Construction Learning

Studies examining the effect of a high-frequency exemplar verb in the acquisition of novel constructions often do so in the context of an artificial language learning task, where children are introduced to a novel verb mapped to a particular event semantics (Casenhiser and Goldberg, 2005; Wonnacott et al., 2008). We use our computational model to investigate the existence and role of an exemplar verb in the acquisition of the English finite SC syntax — a complex structure that has received less attention in such experimental studies (though see Kidd et al. (2006, 2010)). The psycholinguistic studies mentioned above involve children, and hence are often limited in the number of items they can investigate, and in how much they can tease apart the various interacting factors that might play a role in the results (e.g., Casenhiser and Goldberg (2005); Kidd et al. (2006); Wonnacott et al. (2008); Kidd et al. (2010)). Importantly, the use of a computational model enables us to vary distributional properties of the input in a way not easily achieved in a human experimental setting.

For example, the sentence recall tasks performed by Kidd et al. (2010) examine only eight verbs that can occur with the finite SC syntax (i.e., finite-complement-taking (FCT) verbs). Moreover, due to their choice of verbs, they cannot separate the effects in their results of overall verb frequency and relative frequency of use of finite-SC with the Belief construction. Using a computational model of argument structure learning, we extend these investigations into a larger set of FCT verbs, and also manipulate the input in such a way that we can tease apart the effects of the various frequency factors. Our results are consistent with the findings of Kidd et al. (2006, 2010), that the relative frequency of a verb with a sentential complement is positively correlated with the ability of young

---

1Following Kidd et al. (2010) we focus here on verb that occur with the finite SC, as opposed to other SC types, such as infinitival SC. This group includes MSVs, such as think and know, and non-MSVs, such as say and see. We refer to this group as finite-complement-taking (FCT) verbs parallel to the term complement-taking verbs (CTVs) used by Kidd et al. (2010).
children to recall sentences containing the verb in that construction. However, through computational modelling, we are further able to provide evidence regarding the possible interaction of verb frequency and relative construction frequency in accounting for their findings.

We aim to verify the role of each of the distributional properties of the emergent exemplar verb in our simulation above in an additional set of experiments. Inspired by the work of Casenhiser and Goldberg (2005), we study the role of a high-frequency exemplar verb (think) in the acquisition of finite-SC, but we do so in the context of a diverse set of verbs and constructions, as is the case in the naturalistic input that children receive. Our results suggest that the acquisition of a construction is facilitated by the relative frequency with which a class of semantically-related verbs appear with the syntactic form associated with the construction. This analysis sheds light on the enabling factors in the acquisition of Belief verbs that are associated with finite-SC and the hard-to-observe belief meaning.

4.3.1 The Emergence of an Exemplar Verb in Construction Learning

Our goal here is to examine the role of verbs’ overall frequency and their frequency with finite-SC, and the interaction of these frequencies, in the acquisition of argument structure constructions. Our simulations are inspired by the imitation task in which participants are asked to repeat a recently-heard utterance. Kidd et al. (2006, 2010) use this approach to examine the effect of verb frequency with finite-SC on how well children repeat utterances, in particular focusing on the relation between frequency of a verb with finite-SC and its likelihood of being correctly repeated, or substituted by another verb.
Table 4.2: The 13 FCT verbs in our data, along with their overall frequency in the data and their relative frequency with finite-SC. Verbs are grouped by semantic class, and only the 13 verbs that appear with this construction are listed.

### Experimental Setup

Following Kidd et al. (2006, 2010), we focus on whether our model correctly repeats the verb of a sentence in an imitation task involving FCT verbs presented with SCs. Notice that following the definition of Kidd et al. for FCT verbs, this group of verbs includes any verb that occurs with finite-SC in our data, i.e., not only MSVs. We present the model with a full input frame representing a complete utterance plus its corresponding scene, analogous to the presentation of a sentence with an accompanying picture, as in the psycholinguistic experiments. We then ask the model to predict the best verb in response to that frame, essentially asking it to repeat the just-presented verb.

To consider the responses of the model over a developmental trajectory, we train it on the full corpus, and at periodic points during training, we present it with a test frame for each of the 13 FCT verbs in our lexicon, to see how it responds to each FCT verb. All the test frames have the same syntactic features (i.e., *syntactic pattern*, *argument count*, and *complement type*) corresponding to a finite-SC that contains a transitive action verb, paired with the appropriate semantic features for the given FCT verb (see Table 4.2). For
consistency, we use the same physical action verb at every test point for the embedded verb (*other predicate*) in all 13 test frames, but randomly vary this verb across each of 100 simulations.

As do Kidd et al. (2006, 2010), we focus on the patterns of *verb repetition* and *verb substitution* among the model’s responses. We record for each of the 13 test frames (at each point of testing) which verb the model predicts as its best response to that frame. To do this, we calculate the likelihood of each of our 31 verbs $v$ given a test frame $F_{\text{test}}$, as in:

$$
P(v|F_{\text{test}}) = \sum_{k \in \text{Clusters}} P_{\text{main}}(v|k)P(k|F_{\text{test}})$$

where $P_{\text{main}}(v|k)$ is the probability of the main predicate feature having the value $v$ in cluster $k$, calculated as in Equation (3.6) on page 44, and $P(k|F_{\text{test}})$ is calculated as in Equation (3.2) on page 44 (see Section 3.2 for details). The model’s response is taken to be the verb with the highest likelihood; this resembles the single choice of a verb made by the participants in the psycholinguistic experiments.

**Results: Verb Repetition**

Kidd et al. (2006, 2010) observe a positive correlation between the frequency of a verb with finite-SC and the proportion of its correct repetitions. We first focus on how frequency with finite-SC impacts our model’s correct repetition of a verb. Figure 4.1 presents the proportion of times that each of the 13 FCT verbs are correctly repeated, which we refer to as the *repetition accuracy*. According to these results, a high frequency with finite-SC is neither a necessary nor a sufficient condition for a verb to be correctly repeated by our model. For example, although the two verbs with the highest repetition accuracy (i.e., *think* and *say*) have high frequencies with finite-SC, other verbs with high frequency with finite-SC (i.e., *bet, guess, know, wish*, and *hope*) are not easy for our model.
to repeat. In addition, see is among the top four verbs to be correctly repeated, although it has relatively low frequency with finite-SC (see Table 4.2). These results suggest that other factors beyond the frequency with finite-SC examined by Kidd et al. (2006, 2010) may play a role here.

Our model enables us to explore some of the possible factors that may affect the repetition accuracy of a verb together with its frequency with finite-SC. The analysis of the possible factors can later be verified through experiments with children. We turn to focus on how the overall frequency of the verb affects the model’s responses: of the four highest-frequency verbs (think, see, say, know), three also have a higher repetition accuracy compared with the other FCT verbs. However, like frequency with finite-SC, overall frequency alone does not predict the responses: the repetition rate is not in frequency order, and know is high frequency but has a low repetition rate. In fact, we note that, except for the verb think, the model rarely repeats Belief verbs correctly, regardless of their frequencies. These results point to another factor that might affect the performance of our model in repeating a verb: the frequency with which semantically-related verbs appear with the same syntactic pattern as the verb to be repeated. Semantically-related verbs such as know and think are likely to occur in the same clusters based on usages with similar semantic and syntactic properties. However, the higher overall frequency of think compared to know is likely to result in a higher number of occurrences of think in such clusters. Therefore, when applying Equation 4.1 to determine the likelihood of a verb, think would obtain higher repetition accuracy based both on its high overall verb frequency that leads to a higher number of usages, and its high frequency with the finite-SC pattern that leads to frequent usage with this pattern. To further understand the interaction of overall frequency and frequency with finite-SC, and distribution over semantic classes, we next look at the patterns of verb substitution by our model.
Interestingly, Kidd et al. (2006, 2010) found that in a large number of cases, children specifically substituted the verb *think* in place of the verb they heard. They thus suggest that *think* is an ‘exemplar’ for the finite-SC construction. Figure 4.2 presents the proportion of times each of the 13 FCT verbs is produced by our model in place of the other verbs out of the model predictions of a different verb than the one presented in the test frame. We refer this rate as the *substitution rate*. That is, for each verb \( v \), its substitution rate reflects the proportion of times that our model incorrectly produces \( v \) in response to the test frames for the other 12 FCT verbs out of its response to each test frame over the 100 simulations. In line with the findings of Kidd et al. (2006, 2010), the model substitutes the verb *think* for the other verbs with a very high likelihood from an early stage.

Kidd et al. (2006, 2010) attribute their finding to the high frequency of the verb *think* with finite-SC. However, we have observed that *think* also has the highest overall frequency among the 13 FCT verbs (see Table 4.2). In addition, *think* is a Belief verb, and it is known that people form a strong association between Belief verbs and the finite-
SC syntax (Gleitman et al., 2005). It is thus not clear whether the status of think as an exemplar for the finite-SC construction is solely due to its high frequency with finite-SC, or if it is also affected by these other factors: (a) the high overall frequency of think, and/or (b) the overall strong connection of Belief verbs to the construction. We explore these factors in the next set of experiments.

4.3.2 Interaction of the Distributional Properties

One of the advantages of using a computational model is that we can manipulate the input to study the effects and interactions of the different frequency factors. Here, we manipulate the input such that we can examine the effects on the substitution patterns in our model of: overall verb frequency, frequency with finite-SC, as well as the frequency with finite-SC of a verb class as a whole. As noted above, think conflates all of these factors: it has high verb frequency, it is frequent with finite-SC, and it belongs to the Belief verb class, which is frequent with finite-SC.

To tease apart these factors in verb substitution, we perform three new experiments, in each of which we change the overall verb frequency of think to be of lower frequency.
We aim to analyze whether *think* would gain the highest substitution rate even when it does not hold the same high overall verb frequency. Moreover, to analyze whether the high overall verb frequency is enough to become the verb with the highest substitution rate, in each of these three experiments we give the high overall frequency *think* originally had to another verb. As described in Section 3.3, we generate our input automatically based on the overall verb frequency of each verb (see Table 4.2 for the overall frequencies of each of the FCT verbs). Verbs with higher overall verb frequency would have proportionally higher number of occurrences in the input data for each simulation. Therefore, by switching the overall frequencies of *think* with another verb, *v*, *think* would have fewer occurrence in the input while *v* would have higher number of occurrences as *think* had before the switch.

To tease apart the three factors under investigation, we perform these experiments by picking a verb for each of the three experiments that either has lower overall verb frequency, lower frequency with finite-SC, or a verb that belongs to a different verb class, e.g., a Communication verb. For this purpose we choose the following three verbs: (i) *guess* - a Belief verb with lower overall verb frequency, (ii) *believe* - a Belief verb with lower frequency with finite-SC, and (iii) *tell* - a Communication verb (see Table 4.2 for a full list of overall frequency and relative frequency with finite-SC for all the verbs in our lexicon). This method allows us to make minimal changes to the overall distributional properties of the verbs in the input by manipulating only two verbs at a time. In practice, we switch in our input generation lexicon the values of two cells in the column “Overall frequency” in Table 4.2 (see Table 3.3 on page 50 for an example of an entry in the input generation lexicon).

Initially, we run 100 additional simulations with *guess* as the most frequent Belief verb instead of *think*, while *think* holds the observed overall frequency of *guess* (i.e., *guess* now holds a frequency of 13829 occurrences, and *think* holds a frequency of 278 occurrences). Importantly, in each set of such simulations we switch the overall frequency
of exactly two verbs while the remaining 29 verbs hold the overall frequency observed in CDS (see Section 3.3 for detailed description of the use of overall frequencies in our input generation process). By making this alteration, the modified most-frequent Belief verb, *guess*, now has a lower relative frequency with finite-SC of 76% compared with the 100% of *think*. We find that, as in the original results with *think*, *guess* is substituted for other verbs a very high proportion of the time (75%) only slightly less than for *think* when it is most frequent.

We next run 100 additional simulations with *believe* as the most frequent Belief verb instead of *think* (i.e., *believe* now holds a frequency of 13829 occurrences, and *think* holds a frequency of 78 occurrences). Given these experimental settings, *believe* is the most frequent Belief verb with a relative frequency with finite-SC of just 21%. This explores the impact of a relatively low frequency with finite-SC in the context of a very high overall frequency, but still within the same semantic group of verbs, i.e., MSVs. As opposed to the previous set of simulations, when *believe* is the highest-frequency verb, the Belief verb with next highest overall frequency (of 7198 occurrences in the input) and relatively high frequency with finite-SC (*know*) (i.e., 61%) becomes the verb most often substituted for others, with a substitution rate of 43%. This behaviour predicts that both a high overall frequency and a relatively high frequency with finite-SC are required for a verb to be treated as an ‘exemplar’ of the finite-SC construction.

We also examine the result of making *tell*, which is not a Belief verb, the highest frequency verb with finite-SC, *think* (i.e., *tell* now holds a frequency of 13829 occurrences, and *think* holds a frequency of 2953 occurrences). We chose *tell* since it has the highest frequency with finite-SC (64%) among the non-MSVs (i.e., Action, Perception, and Communication verbs). Interestingly, although *tell* is a verb with a relatively high frequency with finite-SC (like *guess* above), *tell* does not become the verb the model most frequently substitutes for other verbs (in contrast to *guess*). In this case, *know* — a Belief verb — is the verb most frequently substituted for others. This suggests that
the semantics of the verb also plays an important role in determining the substitution behaviour. The strong association of particular (Belief-verb) semantics with the finite-SC syntactic pattern is necessary to the verb substitution behaviour. Given the switch of the overall frequencies of tell and think, know becomes the most frequently substituting verb as the verb with the highest overall frequency and high frequency with finite-SC among the Belief verbs.

### 4.3.3 The Role of Skewed Distribution in Learning Constructions

The experiments in this section further examine the role of verb, construction, and semantic verb class frequencies in the acquisition of the finite-SC. Above we replicated the emergence of an exemplar verb for the finite-SC syntax. In this section we focus on how the distributional properties that lead to the emergence of an exemplar verb may affect the learning of the associated construction, i.e., the ‘Belief–SC’ construction.

#### Experimental Setup

Following Casenhiser and Goldberg (2005), we focus here on the effect of the distributional pattern of verb usages with a particular construction on the acquisition of that construction. Casenhiser and Goldberg (2005) introduce a novel syntactic construction (i.e., ‘NP NP V’) representing a novel meaning (i.e., ‘NP appearing suddenly in the second NP’) to 5-to 7-year-old children. The construction is introduced in a training stage by presenting five novel verbs appearing in the construction in two input conditions: the skewed condition where one verb accounts for half of the occurrences of the construction, and the balanced condition with roughly equal number of usages of each verb (see Table 4.3 (a) and (b) for an illustration of the two input conditions). The ability of the participants to generalize the novel meaning to a novel verb in each condition was tested using the preferential-looking paradigm. The participants were presented
with additional novel verbs in the learned construction while watching two candidate videos simultaneously. The participants were asked to touch the screen that displayed the video with the corresponding meaning. The generalization ability of the participants was evaluated as the number of times they choose the appropriate video that displayed the novel meaning (i.e., ‘NP appearing suddenly’) for the novel construction. Participants in the skewed condition were significantly better at generalizing the newly-learned construction to a new novel verb, compared to those in the balanced condition. Casen-
hiser and Goldberg (2005) concluded that the skewed distribution with one verb having relatively high frequency with the construction has a facilitating role in the acquisition of this construction.

Our results in Section 4.3.1 suggest that, in addition to the frequency with finite-SC of the individual verbs, investigated by Casenhiser and Goldberg, their semantic class also influences the learning and use of verbs in a construction. This interaction of semantic classes is not addressed by the artificial language experiment of Casenhiser and Goldberg (2005), since it includes only a single class of verbs (novel verbs in the novel construction). Using a computational model enables us to explore the impact of a skewed vs. balanced distribution in a naturalistic setting. In naturalistic input, verbs that occur in the syntactic frame under investigation (here, the finite SC) range over several semantic classes. In addition, all verbs, including those that occur with this syntactic frame, occur with a variety of additional syntactic frames. We examine how strongly our model learns the Belief–SC construction given the skewed input of our CDS-based data compared to a balanced input we create artificially to follow the experimental settings of Casenhiser and Goldberg (2005).

Our artificially-crafted balanced input maintains some of the naturalistic properties of the original skewed distribution, e.g., the same variety of constructions, the same total exposure to FCT verbs, and overall verb frequencies that range from low to high frequency verbs. We achieve this by changing the overall frequency of each of the FCT verbs so that their frequency with finite-SC would be equal (see Table 4.3 (d)). Using our input generation methodology (explained in Section 3.3), in our artificially balanced input each FCT verb would have an equal number of occurrences with finite-SC in each generated input, but these verbs would differ in the number of times they occur in the input. Specifically, we calculate the overall frequency of finite-SC in our data and deduce the average frequency of finite-SC for 13 FCT verbs included in the data. We adjust the overall frequency of each FCT verb in accordance with its relative frequency with
Table 4.3: The 13 FCT verbs in our data, along with their overall verb frequency and their relative frequency with finite-SC in each of the two distributions: originally skewed, and artificially balanced. Verbs are grouped by semantic class, and only the 13 verbs that appear with this construction are listed. The artificially balanced distribution differs from the observed skewed distribution only in the overall frequency of each verb, while the relative frequency with SC remains the same as the skewed distribution.

Importantly, each FCT verb maintains the same variety of occurrences with a range of syntactic frames and has the same relative frequency with finite-SC in both distributions. Note that all FCT verbs have an equal number of occurrences with finite-SC in the balanced input; however, because there is a different number of FCT verbs in each semantic class, the total number of finite-SC usages still slightly differs across classes.
For example, the Belief class contains 5 FCT verbs while the Desire class contains 2 FCT verbs. This difference results, for instance, in a higher frequency with finite-SC for the Belief class compared with the Desire class.

We need to evaluate the ability of the model to generalize its knowledge of the Belief–SC construction in response to a novel verb when training on these two types of input. Casenhiser and Goldberg (2005) engage the participants in preferential looking to evaluate the strength of association to each semantic interpretation. We simulate preferential looking in our model as a choice between possible sets of event primitives that represent the scenes presented to the participants in the preferential looking paradigm. Following the psycholinguistic settings, we construct the test frame with a novel verb in place of the main predicate, where event participants are associated with a belief event, but the semantics of the predicate is missing. In other words, the test frame represents the construction used to test the children, and each set of event primitives represents a candidate test scene in a preferential looking task. We evaluate the likelihood of each set of event primitives included in our data as a possible semantic interpretation of the novel verb. Our evaluation takes into account a wider range of possible interpretations, compared with the use of only two possible meanings by Casenhiser and Goldberg (2005). At each point of testing, over 100 simulations, we record the set of event primitives that the model predicts as its best response to the partial test frame. This prediction corresponds to the selection of the scene with the appropriate action, given the arguments and syntax of the construction (as in Casenhiser and Goldberg (2005)).

Results

Figure 4.4(a) and (b) report the proportion of times given the skewed and balanced input, respectively, that each of the three most likely sets of event primitives is chosen by our model as the most appropriate one, which we refer to as the event prediction rate.\(^2\)

\(^2\)Other sets of event primitives have lower likelihoods than the likelihoods presented here throughout the training stage (i.e., sets that correspond to Action and Desire events).
Figure 4.4: The 3 highest likelihood values of semantic properties of the event (a) given the CDS-based distribution, (b) given artificially balanced frequencies with finite-SC.

Figure 4.4(a) shows that the semantics of Belief events is highly associated with the arguments and syntax of novel Belief verbs from an early stage, given the skewed condition. That is, the Belief-SC construction is strongly entrenched given the naturalistically-skewed input.

However, in the balanced condition, as shown in Figure 4.4(b), only much later in training is the Belief event semantics predicted with the highest rate for the test frames. As in the results presented in Section 4.3.1, there is an effect of overall frequency in addition to frequency with finite-SC, of both verbs and classes. In the balanced input, each FCT verb has the same number of occurrences with finite-SC; hence there is only a small difference in the total number of occurrences of the different classes with this pattern. Recall that, to balance the input in terms of the FCT verb usages, we had to change the overall frequencies of the verbs and classes. For example, the overall frequency of high frequency verbs, e.g., *think*, with high relative frequency with finite-SC, e.g., 100% for *think*, had to be significantly decreased to allow the verb to occur a lower number of times with finite-SC even though most of its occurrences are with this syntactic frame. Therefore, the overall frequency of most Belief verbs has been decreased to fit the balanced distribution. As opposed to the Belief class, most of the Perception verbs had low relative frequency with finite-SC and as a result their overall frequency
in the balanced distribution has been increased accordingly. Given that, we note that the overall frequency of the Belief class in the balanced input is much lower than that of the Perception class. The model is thus exposed to many more usages of Perception verbs than to usages of Belief verbs initially. In early training stages, the model observes more usages of Perception verbs than usages of Belief verbs. Some of these usages would be with the finite-SC, causing the observed delay in the formation of a strong Belief–SC construction. The model requires more input to observe enough usages of Belief verbs to account for the higher relative frequency of the Belief class with finite-SC compared with the Perception class.

4.4 Modelling the Acquisition of Desires before Beliefs

Our results in Section 4.3 suggest that exemplar-based learning of a construction (such as the finite-SC) is sensitive to several properties of the input, including the overall verb frequency, the frequency of each verb with the construction, and the frequency of each semantic verb class with the construction. We next further our investigation to analyze how these distributional properties play a role in the process of learning the association of a syntactic frame with the meaning conveyed by the semantic verb class it frequently occurs with (here Belief verbs and finite-SC). For this purpose, we analyze the acquisition of MSVs while considering in more details the various meanings and syntactic properties of verb classes included in the group of MSVs.

Despite some shared properties, MSVs are a heterogeneous group, with different types of verbs exhibiting different developmental patterns. Specifically, a wealth of research shows that children produce Desire verbs, such as want and wish, earlier than Belief verbs, such as think and know (Shatz et al., 1983; Bartsch and Wellman, 1995; Asplin, 2002; Perner et al., 2003; de Villiers, 2005; Papafragou et al., 2007; Pascual et al., 2008).
Some explanations for this pattern posit that differences in the syntactic usages of Desire and Belief verbs underlie the observed developmental lag of the latter (de Villiers, 2005; Pascual et al., 2008). In particular, Desire verbs occur mostly with an infinitival SC (as in *I want (her) to leave*), while Belief verbs occur mostly with a finite SC (a full tensed embedded clause, as in *I think (that) she left*). Notably, infinitivals appear earlier than finite SCs in the speech of young children (Bloom et al., 1984, 1989). Others suggest that Desire verbs are conceptually simpler (Bartsch and Wellman, 1995) or pragmatically/communicatively more salient (Perner, 1988; Fodor, 1992; Perner et al., 2003). Proponents of the conceptual and pragmatic accounts argue that syntax alone cannot explain the delay in the acquisition of Belief verbs, because children use finite SCs with verbs of Communication (e.g., *say*) and Perception (e.g., *see*) long before they use them with Belief verbs (Bartsch and Wellman, 1995).

We use our computational model of verb argument structure acquisition to shed light on the factors that might be responsible for the developmental gap between Desire and Belief verbs. Importantly, our model exhibits the observed pattern of learning Desire before Belief verbs, without having to encode any differences in difficulty between the two classes in terms of their syntactic or conceptual/pragmatic requirements. The behaviour of the model can thus be attributed to its probabilistic learning mechanisms in conjunction with the distributional properties of the input. In particular, we investigate how the model’s learning mechanism interacts with the distributions of several classes of verbs — including Belief, Desire, Perception, Communication, and Action — in the finite and infinitival SC syntax to produce the observed pattern of acquisition of Desire and Belief verbs. Using a computational model can reveal the potential effects of interactions of verb classes in human language acquisition which would be difficult to investigate experimentally. Our results suggest that the distributional properties of the relevant verb classes are a potentially important, and heretofore unexplored, factor in experimental studies of the developmental lag of Belief verbs.
4.4.1 Experimental Setup

<table>
<thead>
<tr>
<th>Semantic class</th>
<th>Verb</th>
<th>Frequency</th>
<th>% Relative frequency with SC-fin</th>
<th>SC-inf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belief</td>
<td>think</td>
<td>13829</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>bet</td>
<td>391</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>guess</td>
<td>278</td>
<td>76</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>know</td>
<td>7189</td>
<td>61</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>believe</td>
<td>78</td>
<td>21</td>
<td>-</td>
</tr>
<tr>
<td>Desire</td>
<td>wish</td>
<td>132</td>
<td>94</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>hope</td>
<td>290</td>
<td>86</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>want</td>
<td>8425</td>
<td>-</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>like</td>
<td>6944</td>
<td>-</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>need</td>
<td>1690</td>
<td>-</td>
<td>60</td>
</tr>
<tr>
<td>Communication</td>
<td>tell</td>
<td>2953</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>say</td>
<td>8622</td>
<td>60</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ask</td>
<td>818</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>speak</td>
<td>62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>talk</td>
<td>1322</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Perception</td>
<td>hear</td>
<td>1370</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>see</td>
<td>9717</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>look</td>
<td>5856</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>watch</td>
<td>1045</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>listen</td>
<td>413</td>
<td>33</td>
<td>2</td>
</tr>
<tr>
<td>Action</td>
<td>go</td>
<td>20364</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>get</td>
<td>16493</td>
<td>-</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>make</td>
<td>4165</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>put</td>
<td>8794</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>come</td>
<td>6083</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>eat</td>
<td>3894</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>take</td>
<td>3239</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>play</td>
<td>2565</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sit</td>
<td>2462</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>give</td>
<td>2341</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>fall</td>
<td>1555</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4.4: The list of our 31 verbs from the five semantic classes, along with their overall frequency, and their relative frequency with the finite SC (SC-fin) or the infinitival SC (SC-inf).

Psycholinguistic studies have used variations of a novel verb prediction task to examine how strongly children (or adults) have learned to associate the various syntactic and semantic properties of a typical MSV usage (see Section 3.1.1 for a detailed description of the novel verb prediction task). In particular, the typical Desire verb usage combines desire semantics with an infinitival SC syntax, while the typical Belief verb usage combines
belief semantics with a finite SC syntax. In investigating the salience of these associations in human experiments, participants are presented with an utterance containing a nonce verb with an SC (e.g., *He gorp d that his grandmother was in the bed*), sometimes paired with a corresponding scene representing a mental event (e.g., a picture or a silent video depicting a *thinking* event with heightened saliency). An experimenter then asks each participant what the nonce verb (*gorp*) “means” — i.e., what existing English verb does it correspond to (see, e.g., Asplin, 2002; Papafragou et al., 2007). The expectation is that, for example, if a participant has a well-entrenched belief construction, then they should have a strong association between the finite-SC syntax and belief semantics, and hence should produce more Belief verbs as the meaning of a novel verb in a finite-SC utterance (and analogously for infinitival SCs and Desire verbs).

We perform simulations that are based on such psycholinguistic experiments. After
training the model on some number of input frames, we then present it with a test frame in which the main verb (head predicate) is replaced by a nonce verb like *gorp* (a verb that does not occur in our lexicon). Analogously to the human experiments, in order to study the differences in the strength of association between the syntax and semantics of Desire and Belief verbs, we present the model with two types of test frames: (i) a typical desire test frame, with syntactic features corresponding to the infinitival SC syntax, optionally paired (depending on the experiment) with semantic features associated with a Desire verb in our lexicon; and (ii) a typical belief test frame, with syntactic features corresponding to the finite SC syntax, optionally paired with semantic features from a Belief verb.\(^3\) We measure the likelihood of the model producing a verb from each of the semantic classes following the same methodology presented in Section 3.3.2. Note that here, the Class in Equation (3.10) ranges over the 5 classes in Table 4.4.

### 4.4.2 Experimental Results

The novel verb prediction experiments described above have found differences in the performance of children across the two MSV classes (e.g., Asplin, 2002; Papafragou et al., 2007). For example, children performed better at predicting that a novel verb is a Desire verb in a typical desire context (infinitival-SC utterance paired with a desire scene), compared to their performance at identifying a novel verb as a Belief verb in a typical belief context (finite-SC utterance accompanied by a belief scene). In Section 4.4.2, we examine whether the model exhibits this behaviour in our verb class prediction task, thereby mimicking children’s lag in facility with Belief verbs compared to Desire verbs.

Recall that some researchers attribute the above-mentioned developmental gap to the conceptual and pragmatic differences between the two MSV classes, whereas others suggest it is due to a difference in the syntactic requirements of the two classes. As noted in

\(^3\)Table 4.4 shows that, in our data, Belief verbs occur exclusively with finite clauses in an SC usage. Although Desire verbs occur in both SC-inf and SC-fin usages, the former outnumber the latter by almost 30 to 1 over all Desire verbs.
Chapter 4. The Facilitative Role of Distributional Properties of MSVs

(a) Human participants in Papafragou et al. (2007)

Figure 4.6: (a) Percent verb types produced by adult and child participants given a desire or belief utterance and scene. (b) The model’s verb class likelihoods given a desire or belief test frame. Child stage is represented by 500 input frames compared to the 10,000 input frames for Adult stage.

Section 4.2, we have tailored our representation of Desire and Belief verbs to not build in any differences in the ease or difficulty of acquiring their syntactic or semantic properties. Moreover, the possibility in the model for “misinterpretation” of mental content as action semantics (see Section 3.2) also applies equally to both types of verbs. Thus, any observed performance gap in the model reflects an interaction between its processing approach and the distributional properties of CDS. To better understand the role of the input, in Section 4.4.2 we examine how the distributional pattern of appearances of various semantic classes of verbs (including Belief, Desire, Communication, Perception and Action verbs) with the finite and infinitival SC constructions affects the learning of the two types of MSVs.

Verb Prediction Simulations

Here we compare the verb prediction responses of the participants in the experiments of Papafragou et al. (2007), with those of the model when presented with a novel verb in a typical desire or belief test frame. (See Section 4.4.1 for how we construct these frames.) Papafragou et al. (2007) report verb responses for the novel verb meaning as desire, belief, or action, where the latter category contains all other verb responses. Looking closely at
Chapter 4. The Facilitative Role of Distributional Properties of MSVs

the latter category as reported by Papafragou et al. (2007), we find that most verbs are what we have termed (physical) Action verbs. We thus report the verb class likelihoods of the model for the Belief, Desire, and Action verbs in our lexicon. To compare the model’s responses with those of the children and adults in the study by Papafragou et al. (2007), we report the responses of the model to the test frames at two test points that simulate the observation from children and adults in the above study. We present the results after training the model with 500 input frames as the most adequately resembling the “Child stage” as presented in the psycholinguistic study. We represent the “Adult stage” as the results we obtain from our model after presenting the model with 10,000 input frames, which were sufficient for the model to converge.

Figure 4.6(a) gives the percent verb types from the results of Papafragou et al. (2007); Figure 4.6(b) presents the results of the model. Similarly to the children in Papafragou et al. (2007), the model at earlier stages of learning (“Child stage”) is better at predicting Desire verbs for a desire test frame (0.56) than it is at predicting Belief verbs for a belief test frame (0.42) — cf. 59% Desire vs. 41% Belief prediction for Papafragou et al. (2007). In addition, as for both the child and adult participants of Papafragou et al. (2007), the model produces more Action verbs in a desire context than in a belief context at both stages.

We note that although the adult participants of (Papafragou et al., 2007) perform well at identifying both Desire and Belief verbs, the model does not identify Belief verbs with the same accuracy as it does Desire verbs, even after processing 10,000 input frames (i.e., the “Adult stage”). In Section 4.4.2, we will see that this is due to the model forming strong associations between the Communication and Perception verbs and the SC-fin usage (the typical syntax of Belief verbs). These associations might be overly strong in our model because of the limited number of verbs and verb classes — an issue we will need to address in the future. We also note that, unlike the results of Papafragou et al.

---

4Based on results presented in Table 4, Page 149 in Papafragou et al. (2007), for the utterance and scene condition.
Chapter 4. The Facilitative Role of Distributional Properties of MSVs

(2007), the model only rarely produces Desire verbs in a Belief context. This also may be due to our choice of Desire verbs, which have extremely few SC-fin usages overall.

To summarize, similarly to children (Asplin, 2002; Papafragou et al., 2007), the model performs better at identifying Desire verbs compared to Belief verbs. Moreover, we replicate the experimental results of Papafragou et al. (2007) without encoding any conceptual or syntactic differences in difficulty between the two types of verbs. Specifically, because the representation of Desire and Belief classes in our experiments does not build in a bias due to the ease of processing Desire verbs, the differential results in the model must be due to the interaction of the different distributional patterns in CDS (see Table 4.4) and the processing approach of the model. Although this finding does not rule out the role of conceptual or syntactic differences between Desire and Belief verbs in delayed acquisition of the latter, it points to the importance of the distributional patterns as a potentially important and relevant factor worth further study in human experiments. We further investigate this hypothesis in the following section.

A Closer Look at the Role of Syntax

We hypothesize that the distribution of the finite-SC and the infinitival-SC syntax across the 31 verbs included in the input to the model enables the replication of the pattern in the above section. The goal of the experiments presented here is to understand how an interaction among the 5 different semantic classes of verbs, in terms of their distribution of appearance with the two types of SC constructions, coupled with the probabilistic “misinterpretation” of MSVs in the model, might play a role in the acquisition of Desire before Belief verbs. Therefore, we aim here to analyze the associations formed by the model between each of the SC syntactic patterns and each of the 5 semantic classes that can occur with these syntactic patterns. Because our focus is on the syntactic properties of the verbs, we present the model with partial test frames containing a novel verb and syntactic features that correspond to either a finite SC usage (the typical use of a Belief
Chapter 4. The Facilitative Role of Distributional Properties of MSVs

Verb prediction given an isolated utterance has been performed with adult participants (e.g., Gleitman et al., 2005; Papafragou et al., 2007). Here we simulate the settings of such experiments, but do not compare our results with the experimental data, since they have not included children.

Figure 4.7: The model’s verb class likelihoods for the individual semantic classes.

(verb) or an infinitival SC usage (the typical use of a Desire verb).\(^5\) We refer to the partial test frames as SC-fin or SC-inf test frames. We test the model periodically, over the course of 10,000 input frames, in order to examine the progression of the verb class likelihoods over time.

First, we examine the verb class prediction likelihoods, given an SC-inf test frame; see Figure 4.7(a). We can see that all through training, the likelihoods are mainly divided between Desire and Action verbs, with the Desire likelihood improving over time. Looking at Table 4.4, we note that the Desire and Action verbs have the highest frequency of occurrence with SC-inf (taking into account both the overall frequency of verbs, and their relative frequency with SC-inf), contributing to their strength of association with the infinitival-SC syntax. Note that the very high likelihood of Action verbs given an SC-inf test frame, especially at the earlier stages of training, cannot be solely due to their occurrence with SC-inf, since these verbs mostly occur with other syntactic patterns. Recall that the model incorporates a mechanism that simulates a higher probability of erroneously attending to the physical action (as opposed to the mental event) at earlier stages, simulating what has been observed in young children (see Section 3.2 for details). We believe that this mechanism is responsible for some of the Action verb responses of

\(^5\)
Next, we look at the pattern of verb class likelihoods given an SC-fin test frame; see Figure 4.7(b). We can see that the likelihoods here are divided across a larger number of classes — namely, Action, Communication, and Perception — compared with Figure 4.7(a) for the SC-inf test frame. Since Action verbs do not occur in our data with SC-fin (see Table 4.4), their likelihood here comes from the misinterpretation of mental events (accompanied with SC-fin) as action. The initially high likelihoods of Communication and Perception verbs result from their high frequency of occurrence with SC-fin. Because at this stage Belief verbs are not always correctly associated with SC-fin due to the high probability of misinterpreting them as action, we see a lower likelihood of predicting Belief verbs. Eventually, the model produces more Belief responses than any other verb class, since Beliefs have the highest frequency of occurrence with the finite-SC syntax.

To summarize, our results here confirm our hypothesis that the distributional properties of the verb classes with the finite and infinitival SC patterns, coupled with the learning mechanisms of the model, account for the observed developmental pattern of MSV acquisition in our model. Our results suggest that the distributional patterns play a role in the difference in the learning trajectory of the association between finite-SC and the Belief meaning and infinitival-SC and the Desire meaning. While our first experiment replicated the pattern of delayed association, this experiment provided an analysis of the properties of the input that may lead to the delay. We further discuss our results in Section 4.5.

4.5 Discussion

Our findings in Section 4.3 suggest an interaction over the following three distributional factors on the acquisition of a construction: (a) the overall frequency of verbs that appear
with that construction, (b) the relative frequency with which each verb appears with the construction, and (c) the overall strong connection of a semantic class of verbs to the construction. The results of the generalization task further illustrate how the emergence of the exemplar verb can affect the development of an entrenched construction as a result of the distributional properties of the input. These results are in line with the psycholinguistic findings (e.g., Naigles and Hoff-Ginsberg (1998); Casenhiser and Goldberg (2005); Wonnacott et al. (2008); Kidd et al. (2006, 2010)). Moreover, they further our understanding of the exemplar-based learning mechanism by providing a broader investigation of the role of each of the above factors in the context of naturalistic input that contains multiple classes of verbs, each appearing with multiple constructions. Our findings signify the importance of considering the interaction of the various distributional factors in the design of psycholinguistic experiments.

In Section 4.4 we use our computational model of verb argument structure learning to shed light on the factors that might underlie the earlier acquisition of Desire verbs (e.g., wish and want) than Belief verbs (e.g., think and know). Although this developmental gap has been noted by many researchers, there are at least two competing theories as to what might be the important factors: differences in the conceptual/pragmatic requirements (e.g., Fodor, 1992; Bartsch and Wellman, 1995; Perner et al., 2003), or differences in the syntactic properties (e.g., de Villiers, 2005; Pascual et al., 2008). Using a computational model, we suggest other factors that may play a role in an explanation of the observed gap, and should be taken into account in experimental studies with human subjects. Our investigation is led by our findings regarding the distributional properties that play a role in the formation of an exemplar verb for a construction.

First, we show that the model exhibits a similar pattern to children, in that it performs better at predicting Desire verbs compared to Belief verbs, given a novel verb paired with typical Desire or Belief syntax and semantics, respectively. This difference in performance suggests that the model forms a strong association between the desire
semantics and the infinitival-SC syntax — one that is formed earlier and is stronger than the association it forms between the belief semantics and the finite-SC syntax. Importantly, the replication of this behaviour in the model does not require an explicit encoding of conceptual/pragmatic differences between Desire and Belief verbs, nor of a difference between the two types of SC syntax (finite and infinitival) with respect to their ease of acquisition. Instead, we find that what is responsible for the model’s behaviour is the distribution of the semantic verb classes (Desire, Belief, Perception, Communication, and Action) with the finite and infinitival SC syntactic patterns in the input.

Children are also found to produce semantically-concrete verbs, such as Communication (e.g., say) and Perception verbs (e.g., see), with the finite SC before they produce (more abstract) Belief verbs with the same syntax. Psycholinguistic theories have different views on what this observation tells us about the delay in the acquisition of Belief verbs. For example, Bartsch and Wellman (1995) suggest that the earlier production of Communication verbs shows that even when children have learned the finite-SC syntax (and use it with more concrete verbs), they lack the required conceptual development to talk about the beliefs of others. Our results suggest a different take on these same findings: because Communication (and Perception) verbs also frequently appear with the finite-SC syntax in the input, the model learns a relatively strong association between each of these semantic classes and the finite SC. This, in turn, causes a delay in the formation of a sufficiently-strong association between the Belief verbs and that same syntax, compared with the association between the Desire verbs and the infinitival SC.

de Villiers (2005) suggests that associating Communication verbs with the finite-SC syntax has a facilitating effect on the acquisition of Belief verbs. In our model, we observe a competition between Communication and Belief verbs, in terms of their association with the finite-SC syntax. To further explore the hypothesis of de Villiers (2005) will require expanding our model to learn to investigate the bootstrapping role of Communication verbs in the acquisition of Beliefs in light of their occurrence with a variety of syntactic
frames as they appear in CDS. The use of SC syntax in the test frame leads to focus on constructions that represent usages with this syntax. However, as Belief and Desire verbs may also occur with non-SC syntax, it is important to analyze how the model forms the association of Belief and Desire meaning given all usages of these semantic properties regardless of the choice of syntactic properties.
Chapter 5

An Incremental Model of Learning Verb Classes

5.1 Introduction

Chapters 3 and 4 analyze the role of SCs in the acquisition of MSVs following the vast psycholinguistic studies of the relation between the two (Shatz et al., 1983; Bartsch and Wellman, 1995; Papafragou et al., 2007). Our simulations in Chapter 3 replicated the difficulty in attending to the mental content given the use of an MSV with an SC. On the other hand, our results in Chapter 4 pointed to the facilitating role of the SC syntax in learning to associate MSVs with their semantic and syntactic properties correctly. In addition, we analyzed the role of the distributional properties over SC types (i.e., finite-SC and infinitival-SC) in the acquisition of MSVs (e.g., Belief and Desire verbs). The goal of this chapter is to expand the analysis of the previous chapters by taking into account the full range of usage of MSVs with a variety of syntactic construction in usages such as: “you believe Danny”, “I want a cookie”, etc.

In contrast to the psycholinguistic attention to the use of SCs with MSVs, another group of studies argues that the complex syntax of SCs cannot be the sole explanatory
factor for the observed delays in the acquisition of MSVs. Such studies provide evidence for their argument from several observations: First, children use finite-SCs with verbs of Communication (e.g., say) and Perception (e.g., see) long before they use them with Belief verbs (Bartsch and Wellman, 1995). The relatively high frequency of these non-MSVs with the SC syntax may be a factor that affects the acquisition of MSVs. For instance, de Villiers and Pyers (2002) and Israel (2008) suggest that children first learn to use the complex SC syntax with conceptually simpler verbs that share aspects of the MSV semantics, e.g., Communication and Perception verbs. This stage might be used to break into the full mental meaning of MSVs using the acquired shared meaning with Communication and Perception verbs. For instance, a child may associate think with expressing one’s point-of-view on a given situation as in “I see daddy is making pancakes”, and only later learn to express abstract mental perception as in “I think it may rain tomorrow”. Secondly, there is evidence that except for very high-frequency MSVs (such as think), many MSVs frequently appear in constructions other than the SC (Diessel and Tomasello, 2001; Klainerman, 2010). Thus, the non-SC usages of MSVs might also play a special role in their pattern of acquisition.

Together, the above studies suggest that in order to capture the full developmental trajectory of MSVs, we need to look at the interaction of the following factors: (i) the overall syntactic behaviour of MSVs, including their appearance with the SC and non-SC syntactic patterns; (ii) the syntactic behaviour of other non-MSV verbs, especially Communication and Perception verbs that have a high-frequency of occurrence with the SC syntax; and (iii) the shared semantic properties of MSVs with other semantic classes of verbs.

We aim to examine the developing interaction of the above three factors, as well as their role in the acquisition of MSVs in their full mental meaning. However, the overall syntactic behaviour, for instance, is usually modeled by several clusters in our model since each of the clusters represents a construction, i.e., the association of verbs over a
common set of syntactic and semantic features. For example, Belief verbs may occur in any of the following constructions: (i) finite SC construction given usages such as “I think daddy is sleeping”, (ii) transitive construction given usages such as “I know him”, (iii) intransitive construction given usages such as “I guess”, etc. These constructions differ in both the syntactic and semantic properties they record, e.g., while a finite SC construction is likely to include many usages with mental meaning, a transitive construction is more likely to include usages such as “I read the news”, and “Danny is watching TV” that share some semantic properties with Belief verbs, but overall differ in their semantic representation. Note that even the semantic properties of the Belief verbs included in each of these constructions may differ, i.e., although all verbs correspond to a mental event, think also has the semantic property of considering while guess has the semantic property of estimating. Therefore, to analyze the association of Belief verbs as a semantic class to other verb classes across syntactic and semantic properties, we require a model that integrates the knowledge learned across all the constructions.

In this chapter, we propose a computational model that extends the existing Bayesian model of verb argument structure acquisition (see Section 3.2) to support the learning of verb classes over the acquired constructions. Compared with the constructions, each verb class captures a higher-level of syntactic and semantic similarity among verb types. The verb classes enable us to expand the analysis in Section 4.4 that focuses on the role of SC syntax in the acquisition of MSVs from the learned constructions. Importantly, our model learn both constructions and verb classes incrementally, which allows us to examine the developmental trajectory as the learned clusters grow over time. Moreover, both types of clustering are monotonic, i.e., we do not re-structure the groupings that our model learns, which allows us to study the formation of the clusters over the course of training. To distinguish the clusters learned in each layer, in this chapter we refer to the clusters learned by the model described in Chapter 3 as constructions, and to the clusters used by the new addition as classes.
To evaluate our verb class model and its design in comparison to previous models that learn verb classes, we first diverge from the analysis of the acquisition of MSVs to explain an additional language acquisition pattern that is typically used to evaluate verb class models (see Section 5.2). In this section, we explain the higher-level generalizations enabled by the verb class model and the specific characteristics of our model in comparison to previous models of verb class learning. We give a review of such computational models in Section 5.2.1. We then present our novel addition to the model of Alishahi and Stevenson (2008) in Section 5.2.2 and compare the performance of the model in the context of the language acquisition pattern studied by previous models. Having validated the ability of our model to learn a useful verb class representation, we present an analysis of the acquisition of MSVs given the verb class knowledge (see Section 5.3). Our results suggest how the properties of CDS may be guiding the acquisition process of MSVs in the interaction of the verbs over a variety of verb classes. Moreover, our results point to the role of non-SC syntax in replicating the gradual association of MSVs to SC syntax and mental meaning.

5.2 Learning the Dative Alternation

Usage-based accounts of language learning note that young children rely on verb-specific knowledge to produce their early utterances (e.g., Tomasello, 2003). There is also evidence that suggests that even young children can generalize their verb knowledge to novel verbs and syntactic frames (e.g., Fisher, 2002), and that the abstract knowledge gradually strengthens over time (e.g., Tomasello and Abbot-Smith, 2002). One area of verb usage where more sophisticated abstraction appears necessary for full adult productivity in language is the knowledge of verb alternations. A verb alternation is a pairing of constructions shared by a number of verbs, in which the two constructions express related argument structures (Levin, 1993): e.g., the dative alternation involves the related forms
of the prepositional dative (PD; \(X \text{ gave } Y \text{ to } Z\)) and the double-object dative (DO; \(X \text{ gave } Z \text{ } Y\)). Such alternations enable language users to readily adapt new and low frequency verbs to appropriate constructions of the language by generalizing the observed use of one such form to the other.

For example, Conwell and Demuth (2007) show that 3-year-old children understand that a novel verb observed only in the DO dative (\(John \text{ gorped } Heather \text{ the book}\)) can also be used in the PD form (\(John \text{ gorped the book to } Heather\)), though the children can only generalize such knowledge under certain experimental conditions. Wonnacott et al. (2008) demonstrate the proficiency of adults in making such generalizations within an artificial language learning scenario, which enables the researchers to explore the distributional properties of the linguistic input that facilitate the acquisition of such generalizations. Their results suggest that the overall frequency of the syntactic patterns as well as the distribution of verbs across the patterns play a facilitatory role in the formation of abstract verb knowledge (in the form of verb alternations) in adult language learners.

In this section, we present our computational model that extends the existing Bayesian model of verb argument structure acquisition of Alishahi and Stevenson (2008) to support the learning of verb classes over the acquired constructions. Importantly, the model can form higher-level generalizations such as learning verb alternations, which is not possible with the model of Alishahi and Stevenson (2008) (cf. the findings of Parisien and Stevenson, 2010). Our model is novel in its approach to verb class formation, because it clusters tokens of a verb that reflect the distribution of the verb over the learned constructions each time the verb is used in an input. That is, the model forms verb classes by clustering verb tokens that reflect the evolving usages of the verbs in various constructions. Moreover, because our model gradually forms its representations of constructions and classes over time (in contrast to other Bayesian models, such as Parisien and Stevenson,

---

The *generalization of an alternation* refers to a speaker using one variant of an alternation for a verb (e.g., PD) having only observed the verb in the other variant (e.g., DO).
2010; Perfors et al., 2010), it is possible to analyze the monotonically-growing representations and show their compatibility with the developmental patterns seen in children (Conwell and Demuth, 2007). We also replicate some of the observations of Wonnacott et al. (2008) on the role of distributional properties of the language in influencing the degree of generalization over an alternation.

5.2.1 Related Work

This section reviews previous computational models of learning verb classes. Specifically, we review the properties of each model and explain how our model addresses the shortcomings of the previous models. For a more detailed review of additional computational models of language acquisition, please refer to Section 2.4.

The hierarchical Bayesian models of Perfors et al. (2010) and Parisien and Stevenson (2010) focus on learning higher-level generalizations about verbs. The model of Perfors et al. (2010) learns verb alternations, i.e., pairs of syntactic patterns shared by certain groups of verbs. By incorporating this sort of abstract knowledge into their model, Perfors et al. (2010) are able to simulate the ability of adults to generalize across verb alternations (as in Wonnacott et al., 2008). That is, Perfors et al. (2010) predict the ability of a novel verb to occur in a syntactic structure after exposure to it in the alternative pattern of that alternation. However, this model is trained on data that contains only a limited number of verbs and syntactic patterns, unlike naturalistic CDS, and moreover incorporates built-in information about verb constructions.

The hierarchical Dirichlet model of Parisien and Stevenson (2010) addresses these limitations by working with CDS data. Moreover, the model of Parisien and Stevenson (2010) simultaneously learns constructions as in Alishahi and Stevenson (2008) and verb classes based on verb alternation behaviour, showing that the latter level of abstraction is necessary to support effective learning of verb alternations. Still, the models of both Parisien and Stevenson (2010) and Perfors et al. (2010) can only be utilized as a batch
process and hence are limited in the analysis of developmental trajectories. Although it is possible to simulate development by training such models on increasing portions of input, such an approach does not ensure that the representations given $n + i$ inputs can be developed from the representation given $n$ inputs. This approach requires several independent simulations given each input size settings. The verb classes obtained from each simulation may not necessarily develop from one another.

In this chapter, we propose a significant extension to the model of Alishahi and Stevenson (2008), by adding an extra level of abstraction that incrementally learns verb classes by drawing on the distribution of verbs over the learned constructions.

### 5.2.2 The Computational Model

The model presented in this section uses a Bayesian clustering algorithm to learn clusters of verb usages that occur in similar argument structure constructions, as in the original model of Alishahi and Stevenson (2008). To this, we add another level of abstraction that learns clusters of verbs that exhibit similar distributional patterns of occurrence across the learned constructions—that is, classes of verbs that occur in similar sets of constructions, and in similar proportions. To distinguish between the clusters of the two levels of abstraction in our new model, we refer to the clusters of verb usages as constructions, and to the groupings of verbs given their distribution over those constructions as verb classes.

**Overview of the Model**

As in our earlier model, our new model learns from a sequence of frames, where each frame is a collection of features representing what the learner might extract from an utterance s/he has heard. Similarly to previous computational studies (e.g., Parisien and Stevenson, 2010), here we focus on syntactic features since our goal is to understand the acquisition of acceptable syntactic structures of verbs independently of their meaning, as in some
relevant psycholinguistic (Wonnacott et al., 2008) and computational studies (Parisien and Stevenson, 2010). Therefore, the representation of the input in this sub-section differs from the earlier experimental settings used in this thesis in using only syntactic features. We focus here particularly on properties such as syntactic slots and argument count. (These features, as in (Parisien and Stevenson, 2010), provide a more flexible and generalizable representation of a syntactic structure than the syntactic pattern string used by Alishahi and Stevenson (2008).) See the bottom rows of boxes in Figure 5.1 for sample input verb usages with their extracted frames.

As described in Section 3.2, the model incrementally clusters the extracted input
frames into constructions that reflect probabilistic associations of the features across similar verb usages; see the middle level of Figure 5.1. Such constructions allow for some degree of generalization over the observed input; e.g., when seeing a novel verb in a Transitive utterance, the model predicts the similarity of this verb to other Action verbs appearing in that pattern (Alishahi and Stevenson, 2008).

Grouping of verb usages into constructions is not sufficient for making higher-level generalizations across verb alternations, as shown by Parisien and Stevenson (2010). Knowledge of alternations is only captured indirectly in constructions (because usages of the same verb can occur in multiple clusters). Following Parisien and Stevenson (2010), we hypothesize that true generalization behaviour requires explicit knowledge that verbs have commonalities in their patterns of occurrence across constructions; this is the basis for verb classes (Levin, 1993; Merlo and Stevenson, 2000; Schulte im Walde and Brew, 2002).

To capture this, our model learns groupings of verbs that have similar distributions across the learned constructions. These groupings form verb classes that provide a higher-level of abstraction over the input; see the top level in Figure 5.1. Consider the dative alternation: the classes capture the fact that some verbs may occur only in prepositional dative (PD) forms, such as sing, while others occur only in double object (DO) forms (call), while still others alternate – i.e., they occur in both (bring).

Our model simultaneously learns both of these types of knowledge: constructions are clusters of verb usages, and classes are clusters of verb distributions over those constructions. The model in both levels is clustering verb tokens – i.e., the features corresponding to the verb at that time in the input, its usage or its current distribution – so that the same verb type may be added to various clusters at different stages in the training.

Because the associations are probabilistic, a linguistic construction may be represented by more than one cluster.
Learning Verb Classes

Our new model learns verb classes by grouping verbs into classes on the basis of their distribution across the learned constructions. That is, verbs that have statistically-similar patterns of occurrence across the learned constructions will be considered as forming a verb class. For example, in Figure 5.1 we see that bring and read may be put into the same class because they both occur in a similar relative frequency across the DO and PD constructions (the leftmost and rightmost constructions in the figure).

We use the same incremental Bayesian clustering algorithm for learning the verb classes as for learning constructions. At the class level, the feature used for determining similarity of items in clustering is the distribution of each verb across the learned constructions. As for constructions, the model learns the verb classes incrementally; the number and type is not predetermined. Moreover, just as constructions are gradually formed from successively processing a particular verb usage at each input step, the model forms verb classes from a sequence of snapshots of the input verb’s distribution over the constructions at each input step. This means that our model is forming classes of verb tokens rather than types; if a verb’s behaviour changes over the duration of the input, subsequent tokens (the distributions over constructions at later points in time) may be clustered into a different class (or classes) than earlier tokens, even though prior decisions cannot be undone.

Formally, after clustering the input frame at time $t$ into a construction, as explained in Section 3.2, the model extracts the current distribution $d_v$ of its head verb $v$ over the learned constructions at time $t$; this is estimated, similarly to Equation (3.6) on page 44, as a smoothed version of $v$’s relative frequency in each construction:

$$P(k|v) = \frac{\text{count}(v, k) + \lambda}{n_v + \lambda \alpha_k}$$

(5.1)

where $\text{count}(v, k)$ is the number of times that inputs with verb $v$ have been clustered
into construction $k$, and $n_v$ is the number of times $v$ has occurred in the input thus far. The parameter $\lambda$ is set to a small value ($\lambda << 1$) to assign a non-zero probability to constructions that have not yet recorded usages of the verb $v$. The use of $\lambda$ here resembles its use in Equation (3.6). The smoothing parameter $\alpha_k$ estimates the number of possible constructions that usages of the verb can be clustered to, similar to the use of $\alpha_i$ in Equation (3.6). $d_v$ is represented as a vector of the probabilities $P(k|v)$ for each construction $k$.

To cluster this snapshot of the verb’s distribution, $d_v$, it is compared to the distributions encoded by the model’s classes. The distribution $d_c$ of an existing class $c$ is the weighted average of the distributions of its member verb tokens:

$$d_c = \frac{1}{|c|} \sum_{v \in c} d_{vc} \times \text{count}(v, c)$$  \hfill (5.2)

where $|c|$ is the size of class $c$, $\text{count}(v, c)$ is the number of occurrences of $v$ that have been assigned to $c$, and $d_{vc}$ is the distribution of the verb $v$ given by the tokens of $v$ in class $c$ (the “snapshots” of distributions of $v$ assigned to class $c$). That is, $d_{vc}$ is an average of the distributions of all $d_v$ for verb $v$ that have been clustered into $c$.

The model finds the best class for a given verb distribution $d_v$ based on its similarity to the distributions of all existing classes and a new one:

$$\text{BestClass}(d_v) = \argmax_{c \in \text{Classes}} (1 - D_{JS}(d_c \parallel d_v))$$  \hfill (5.3)

where $c$ ranges over all existing classes as well as a new class that is represented as a uniform distribution over the existing constructions. Jensen–Shannon divergence, $D_{JS}$, is a popular method for measuring the distance between two distributions: It is based on the KL–divergence, but it is symmetric and has a finite value between 0 and 1:

$$D_{JS}(p \parallel q) = \frac{1}{2} D_{KL}(p \parallel \frac{1}{2}(p + q)) + \frac{1}{2} D_{KL}(q \parallel \frac{1}{2}(p + q))$$  \hfill (5.4)
Table 5.1: Number of non-alternating (non-ALT) and alternating (ALT) verbs in our lexicon, as well as the relative frequency of each construction in our generated input corpora.

<table>
<thead>
<tr>
<th></th>
<th>non-ALT</th>
<th>ALT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DO-only</td>
<td>PD-only</td>
</tr>
<tr>
<td>Number of verbs</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Relative frequency</td>
<td>14%</td>
<td>2%</td>
</tr>
</tbody>
</table>

5.2.3 Experimental Setup

Generation of the Input Corpora

We follow the input generation method of Alishahi and Stevenson (2008) as used in the previous chapters. Our input-generation lexicon contains 71 verbs consisting of the 31 verbs included in our lexicon in previous chapters and an additional 40 verbs of the most frequent verbs in CDS, in order to have a range of verbs that occur with the PD and DO constructions. Table 5.2.3 shows the number of verbs that appear in the DO or PD construction only (non-alternating), as well as those that alternate across the two. (The table also gives the relative frequency of each dative construction in our generated input corpora.)

Simulations

Similar to the experimental settings of the previous chapters, we conduct 100 simulations for each experiment (each using a different input corpus) to avoid any dependency on specific idiosyncratic properties of a single generated corpus. For each simulation, we train our model on an automatically-generated corpus of 15,000 frames, from which the model learns constructions and verb classes. At specified points in the input, we present the model with usages of a novel verb in a DO and/or PD frame, and then test the model’s generalization ability by predicting DO and PD frames given that verb. Since we are interested in the relative likelihoods of the two frames, we report the difference between the log-likelihood of the DO frame and the log-likelihood of the PD frame, i.e., log-likelihood(DO) − log-likelihood(PD).
Specifically, we form a partial frame $F_{\text{test}}$ (containing all usage features except for the verb) that reflects either the PD or the DO syntax, and assess the probability $P(F_{\text{test}}|v)$ for each of these given a novel verb $v$, as in:

$$P(F_{\text{test}}|v) = \sum_{k \in \text{Constructions}} P(F_{\text{test}}|k) P(k|v)$$ (5.5)

where $P(F_{\text{test}}|k)$ is calculated as in Equation (3.5).

We can calculate $P(k|v)$ in two different ways: using only the knowledge in the constructions of the model, and using the knowledge that takes into account the verb classes over the constructions. For model predictions based on the construction level only, we calculate $P(k|v)$ as in Equation (5.1), which is the smoothed relative frequency of the verb $v$ over construction $k$.

Predictions using knowledge of the verb classes will instead determine $P(k|v)$ drawing on the fit of verb $v$ to the various classes (specifically, the similarity of $v$’s distribution over constructions to the distribution encoded in each class), and the likelihood of each construction $k$ for each class $c$ (specifically, the likelihood of $k$ given the distribution over constructions encoded in $c$), as in:

$$P(k|v) \approx \sum_{c \in \text{Classes}} P(k|c) P(c|v)$$ (5.6)

where $P(k|c)$ is the probability of construction $k$ given class $c$’s distribution over constructions ($d_c$); and $P(c|v)$ is the probability of $c$ given verb $v$’s distribution $d_v$ over the constructions (using Jensen-Shannon divergence as in Equation (5.4)).

Due to the different number of clusters in each of the construction and class layers of the model, the likelihoods computed for each will differ in the range of values. For this reason, specific values cannot be directly compared across the layers of the model, rather we must analyze the general trends of the construction-only and class-based results.
5.2.4 Evaluation

In this section we examine whether and how our model generalizes across the two variants of the dative alternation, the double-object dative (DO) and the prepositional dative (PD). To do so, we measure the tendency of the model to produce a novel verb observed in one dative frame either in that same frame, or in the other dative frame (unobserved for that verb). Our goal is to understand the impact of the learned constructions and classes on this generalization behaviour. Following Parisien and Stevenson (2010), we examine three input conditions in which the novel verb occurs: (i) twice with the DO syntax (non-alternating); (ii) twice with the PD syntax (non-alternating); or (iii) once each with DO and PD syntax (alternating) (for the alternating condition, half the simulations have DO first, and half have PD first). We then ask the model to predict the likelihood of producing each dative frame with that verb. Our focus here is on comparing the generalization abilities of the two levels of abstract knowledge in our model: the constructions versus the verb classes.

As a reminder, we use the dative alternation as one example for considering this kind of higher-level generalization behaviour observed in adults and to a lesser extent in children. Moreover, we perform the analysis in the context of naturalistic input that contains many verbs (those that appear in the dative and those that do not), and a variety of constructions, to provide a realistic setting for the task.

Our settings differ from the psycholinguistic studies in the variability of constructions compared with the artificial language used by Wonnacott et al., and in focusing only on the syntactic properties unlike Conwell and Demuth. However, we follow the settings of these studies in analyzing the syntactic properties of a generated utterance given minimal exposure to a novel verb. Therefore, we aim to replicate their general observations by showing that (i) children are limited in their ability to generalize across verb alternations compared with adults, and (ii) the frequency of a construction has a positive correlation with the generalization rate of the construction. We compare our results in this section
Figure 5.2: The difference between the log-likelihood values of the DO and PD frames, given each of the three input conditions: DO only, PD only, and Alternating. Values above zero denote a higher likelihood for the DO frame, and values below zero denote a higher likelihood for the PD frame.

to the simulations conducted by Perfors et al. (2010) and Parisien and Stevenson (2010) in order to validate the compatibility of our model.

Generalization of Learned Knowledge

We examine the generalization patterns of our model when presented with a novel verb in DO/PD forms after being trained on 15,000 inputs, which we compare to the performance of adults in such language tasks. We first consider the case where the model predictions are based solely on the knowledge of constructions. Here we expect the predictions to correspond to the syntactic properties of the two inputs observed for the novel verb, with limited generalization. That is, we expect a non-alternating verb to be much more likely in the observed dative frame, and an alternating verb to be equally likely in both frames. The left hand side of Figure 5.2 presents the differences in log-likelihoods of the predicted DO and PD frames for the novel verb using the construction-based probabilities. The
results confirm our expectation that the knowledge of constructions can support only limited generalization across the variants of an alternation. For the non-alternating conditions, the observed frame is highly favoured, and for the Alternating test scenario, the DO and PD frames have nearly equal likelihoods.

We next turn to using the knowledge of verb classes, which we expect to enable generalizations that correspond to verb alternation behaviour — that is, we expect the model predictions here to reflect the knowledge that verbs that occur in one form of the alternation also often occur in the other form of the alternation. This is possible because the classes in the model encode the distributional patterns of verbs across constructions. In the absence of other factors, we would expect the Alternating condition to again show near equal likelihoods for the two frames, and the two non-alternating conditions to show a slight preference for the observed frame (rather than the strong preference seen in the construction-based predictions), because the unobserved frame is also likely due to the knowledge here of the alternation.

The right hand side of Figure 5.2 presents the difference in the log-likelihoods of the DO and PD frames when using the knowledge encoded in the verb classes. The results are not directly in line with the simple prediction above: The non-alternating (DO-only and PD-only) conditions show a weak preference (as expected) for one frame over another, but both favour the DO frame, as does the Alternating condition. That is, the PD-only and Alternating conditions show a preference for the DO frame that does not follow simply from the knowledge of alternations.

The DO preference in the PD-only and Alternating conditions arises due to distributional factors in the input, related to the frequencies of the constructions reported in Table 5.2.3. First, the DO frame is overall much more likely than the PD frame, causing generalization in the PD-only and Alternating conditions to lean more to that frame. Second, fully 1/3 of the uses of the PD frame in the corpus are with verbs that alternate (i.e., 1% of the corpus are PD frames of alternating PD-DO verbs, out of a total
of 3% of the corpus being PD frames), while only 1/8 of the uses of the DO frame are with alternating rather than non-alternating verbs. Recall that our classes encode the distribution (roughly relative frequency) of the verbs in the class occurring across the different constructions. This means that in our class-based predictions, greater weight will be given to constructions with DO when observing a PD frame than to constructions with PD when observing a DO frame. These results underline the importance of using naturalistic input and considering the impact of various distributional factors on generalization of verb knowledge.

In contrast to the construction-based results, our class-based results conform with the experimental findings of Wonnacott et al. (2008), who show that adult (artificial) language learners robustly generalize a newly-learned verb observed in a single syntactic form by producing it in the alternating syntactic form under certain language conditions. Moreover, we show similar distributional effects to theirs – the overall frequency of the syntactic patterns, as well as the distribution of verbs across those patterns – in the level of preference for one form over another, within the context of our naturalistic data with multiple verbs, constructions, and alternations. These results show that the verb classes in the model are able to capture useful abstract knowledge that is key to understanding the human ability to make high-level generalizations across verb alternations.
Development of Generalizations

Figure 5.3: Difference of log-likelihood values of the DO and PD frames over the course of training for the constructions and the verb classes for each of the 3 test scenarios. Values above zero denote a higher likelihood for the DO frame, and values below zero denote a higher likelihood for the PD frame.

Next, we present the results of our model evaluated throughout the course of training in order to understand the developmental pattern of generalization. We perform the
same construction-based or class-based prediction tasks (the likelihoods of a DO and PD frame), following the same input conditions (a novel verb with two DO frames, two PD frames, or one of each) at given points during the 15,000 inputs. As above, we present the difference in the log-likelihood values of the DO and the PD frames in order to focus on the relative likelihoods of the two frames within each condition of construction-based or class-based predictions.

Figure 5.3(a) presents the results for the DO-only test scenario. As in Section 5.2.4, for both construction-based and class-based predictions there is a higher likelihood for the DO frame throughout the course of training. In contrast, the incremental results for the PD-only test scenario, in Figure 5.3(b), display a developing level of generalization throughout the training stage for the class-based predictions. While the construction-based predictions reflect a much higher likelihood for the PD frame, the results from the verb classes are in favor of the PD frame only initially; after training on 5000 input frames, the likelihood of the DO frame becomes higher for this test scenario. These results indicate that using construction knowledge alone does not enable generalization from the PD frame to the DO frame; in contrast, the verb class knowledge enables the gradual acquisition of generalization ability over the course of training.

Finally, Figure 5.3(c) presents the results for the Alternating test scenario for the two types of predictions. As in Section 5.2.4, both construction-based and class-based predictions have a small preference for the DO frame. In the construction-based predictions, this preference lessens over time to where the likelihoods for DO and PD are almost equal, while the class-based predictions stay relatively constant in their preference for the DO frame. In some ways the construction-based predictions are more expected in response to an apparently alternating verb; however, the class-based predictions show a higher degree of generalization, responding to the higher frequency of the DO frame and the higher association of PD frames with DO alternates. These results again emphasize the importance of further exploring the role of distributional factors on generalization of
verb knowledge in children.

The developmental results presented here are in line with the suggestions of Tomasello (2003) that the productions of younger children follow observed patterns in the input, and only later reflect robust generalizations of their knowledge across verbs. Conwell and Demuth (2007), for example, found evidence of generalization across verb alternations in 3-year-old children, but their production of unobserved forms for a novel verb was very sensitive to the precise context of the experiment and the distributional patterns across the novel verbs. In accord with these observations, the developmental trajectories in our model show that our class-based predictions increase in their degree of generalization over time, and are sensitive to various distributional factors in the input, such as the overall expectation for a frame and the expectation that a verb will alternate.

5.2.5 Discussion

In this section, we analyzed how generalization is supported by each level of learning in our novel computational model: constructions and verb classes. Our results confirm (cf. Parisien and Stevenson, 2010) that a higher-level knowledge of the verb classes is required to replicate the observed patterns of generalization, such as producing a novel verb *gorp* in the prepositional dative pattern after hearing it in the double object dative pattern. In addition, our analysis of the incrementally developing verb classes shows that the generalization knowledge gradually emerges over time, similar to what is observed in children.

The flexibility of input representation of our model enables us to further explore the properties of the input in learning abstract knowledge, following psycholinguistic studies. Our results replicate the findings of Wonnacott et al. (2008) on the role of the distributional properties over the alternating syntactic forms, but in naturalistic settings of many constructions. Our results here validate the compatibility of our model by both replicating psycholinguistic observations and extending the computational analysis
offered by previous models. We next use our novel model to extend our analysis presented in Chapter 4, using the knowledge represented in the verb class level.

5.3 Gradual Acquisition of Mental State Meaning

In the experimental section above, we show that our extension to the model is required to replicate the generalization ability of children. In this section, we aim to further our analysis of the acquisition of MSVs using the generalization ability of the extended model. As described in the introduction to this chapter, our goal here is to investigate how a learner might interpret a usage of an MSV given the knowledge of the distribution of verbs from various classes over a variety semantic and syntactic properties. We describe our experimental setup and results of this investigation in the following sections.

5.3.1 Experimental Setup

Set-up of Simulations

In this set of simulations, we use the input generation lexicon presented in Section 5.2.3, which offers a more comprehensive list of verbs from the variety of semantic classes included in our data. We adapt the frame representation features used in Section 5.2.3 to incorporate the semantic properties required to analyze the acquisition of the mental meaning as in the previous chapters (see Figure 5.4 for an illustration of the features, constructions, and classes that can be learned in this set of simulations). Our goal here is to evaluate how our model interprets a typical usage of an MSV by generalizing over the accumulated knowledge, resembling the verb prediction task used in many psycholinguistic studies of MSV acquisition (e.g., Asplin, 2002; Papafragou et al., 2007). As described in previous chapters, in this task, participants are asked to predict the meaning of a novel verb (e.g., *gorp*) in a given utterance (e.g., “she gorps that daddy is sleeping”) paired with an image/video depicting a mental event. In chapter 4 we replicated the setting of
Figure 5.4: A visual representation of the two layers in the simulation of gradual Acquisition of Mental State Meaning and of the representation of the distribution of verbs over the constructions used to learn the verb classes. The sample constructions present a small subset of the features and semantic properties used in our model. A full frame includes the predicate, syntactic features (syntactic pattern, argument count, preposition, and complementizer) and semantic features (event primitives and event participants). As in Table 5.1 darker shaded squares in the distribution vector represent higher probability of the verb given the corresponding construction. Note that this illustration simplifies the structure of the layers and distribution using a subset of the possible constructions and classes in our model.

this experiment to compare the prediction of verbs by our model and the participants in the study of Papafragou et al. (2007). Here, our goal is to examine more closely how the verb prediction by our model is driven by the learned knowledge in terms of the semantic interpretation the model gives to the verb at the time of prediction. This simulation examines the ability of the model to correctly associated the verb usage to its typical meaning (e.g., a usage of a Desire with infinitival SC to Desire meaning or a usage of a Belief verb with finite SC to Belief meaning). The extended model allows us to also analyze this association given the verb class knowledge that provides a comprehensive knowledge level across usages in various syntactic construction, with and without SCs.
At the same time, we simulate this task to examine the developmental trajectory of MSV acquisition in our model. For this goal, we aim to evaluate whether the test frames are initially associated to a non-mental meaning, following the hypothesis of de Villiers (2005) and Israel (2008). Using the verb class model we hope to shed light on the factors that might be responsible for the observed developmental patterns of Desire and Belief verbs. That is, we aim to analyze our results in respect to the interaction of Desire and Belief verbs with other (conceptually-simpler) verbs over their shared syntactic and semantic properties that might play a role in the acquisition of MSVs (de Villiers and Pyers, 2002; Israel, 2008).

Specifically, we train our model on a randomly generated input corpus of 10,000 input frames, performing periodic tests: At each test point, we present the model with a test frame that represents the psycholinguistic settings of the verb identification task. We then use the model to estimate the likelihood of the event type corresponding to the 5 semantic verb classes of Belief, Desire, Perception, Communication, and Action. These likelihoods represent our model’s interpretation of a test scenario: E.g., the likelihood of the event type Communication given a Belief-finite test frame reflects how likely it is that our model interprets the novel verb in the frame by generalizing the mental meaning as having a Communication semantics; see the following section for how we estimate the event type likelihoods. Our goal here is to evaluate whether a child hearing an utterance such as “I gorp that daddy is sleeping” describing a Belief event, may interpret the meaning of gorp as Communication, e.g., “I say that daddy is sleeping”, based on a formed association between Belief events and Communication meaning at the time of prediction.

We use two kinds of test frames to represent typical usages of MSVs: (i) Desire-inf containing a novel verb, the infinitival-SC syntactic properties, and the semantic properties of a randomly-chosen Desire verb from our lexicon; and (ii) Belief-fin containing a novel verb, the finite-SC syntactic properties, and the semantic properties of a randomly-
chosen Belief verb from our lexicon. We include the Desire-inf test frame since even for Desire verbs, there is still an initial stage when they are produced mostly in non-mental meaning, despite their earlier stage of production compared with Belief verbs (Bartsch and Wellman, 1995).

**Estimating Event Type Likelihoods**

Recall that each verb entry in our lexicon is represented as a collection of features, including a set of event primitives — e.g., the set associated with the Belief verb *think* is \{*state, cogitate, belief, communicate*\}. We estimate each event type likelihood (e.g., Belief likelihood) by averaging over the likelihoods of all *event primitive sets* corresponding to verbs of that class (e.g., all Belief verbs) according to our lexicon.

Formally, we calculate the likelihood of each event primitive set \(S\) given a test frame \(F_{test}\), as in:

\[
P(S|F_{test}) = \sum_{k \in Clusters} P_{\text{event primitives}}(S|k)P(k|F_{test})
\]

where \(P(S|k)\) is the probability of the primitive set \(S\) given construction \(k\), calculated as in Equation (3.6); and \(P(k|F_{test})\) is the probability of assigning the test frame \(F_{test}\) to construction \(k\). Note that only the constructions encode the individual semantic and syntactic features (including the event primitives). Hence, we need to rely on the model’s learned constructions to estimate \(P(S|k)\). However, we can use two ways of estimating \(P(k|F_{test})\): (i) by drawing on the model’s learned knowledge as reflected in the constructions; and (ii) by drawing on the model’s learned knowledge of verb classes. This would help us understand the role of the model’s learned verb classes in the acquisition of MSVs. We calculate the probabilities according to each layer: the construction layer, i.e., \(P_{L1}(k|F_{test})\), and the verb class layer, i.e., \(P_{L2}(k|F_{test})\), as described in Section 5.2.3 (see Equation (5.1) and Equation (5.6)). Notice that in Section 4.4.1 we predicted the likelihood of each verb following the experimental settings of Papafragou et al. (2007). Therefore, we estimate the likelihood of each semantic class based on the likelihood to
predict each of the verbs in the class. However, in this section we aim to focus on
the semantic class itself as a group of verbs to evaluate our results in comparison to
the hypothesis of de Villiers (2005) that the acquisition of MSVs might be enabled by
prior association to alternative semantic properties. We predict the semantic properties
represented by the event primitive sets to evaluate whether our model goes through the
hypothetical acquisition stages to associate a set of semantic and syntactic properties to
a certain meaning as suggested by the psycholinguistic literature.

5.3.2 Experimental Results

We test our model’s knowledge of MSVs (Belief and Desire verbs) by examining the
event type likelihoods (that we estimate as explained above) of each possible mental
interpretation (i.e., Desire and Belief) or non-mental interpretation (i.e., Perception,
Communication, and Action), given each of our two types of test frames. We say that
the model has acquired a solid knowledge of Belief (Desire) verbs if it assigns the highest
likelihood to the Belief (Desire) event type when presented with a Belief-fin (Desire-inf)
test frame. Our goal is to take into consideration any construction that corresponds to
either syntactic or semantic properties that are typical to an MSV usage. Therefore, we
follow the settings of the psycholinguistic task and include the typical syntactic pattern
used with MSVs (i.e., the SC syntax) to make the model rely on constructions associated
with this pattern. Similarly, we include the semantic properties of Belief or Desire in
the test frames, which has the effect of looking into constructions that reflect mental
semantics even when associated with non-SC syntactic patterns. This way, we can study
the role of both the SC and the non-SC syntax in the acquisition of MSVs. In this way,
we consider the role of the semantic similarity between usages such as “I see your point”
and “I believe him”.

To evaluate the role of verb classes in the acquisition of MSVs, we compare the
developmental patterns in our model arising from each of the two layers. In one case,
Analysis based on the Learned Constructions

Figure 5.5(a) presents the likelihood of each event type given the Desire-inf test frame, while Figure 5.5(b) presents the likelihoods given the Belief-fin test frame. As can be seen,
our model acquires Desire verbs almost instantly (from very early stages of training), but exhibits a delay in its acquisition of Belief verbs. This pattern is similar to what has been observed in children in that Desires are acquired earlier than Belief verbs; however, it is lacking the observed initial stage of not producing Desire verbs in their mental meanings. Interestingly, the earlier acquisition of Desire verbs can be attributed to their higher frequency (compared to Belief verbs) of appearing with non-SC syntax, e.g., *I want a cookie, I like apples.* See Figure 5.6 for these different syntactic distributions of Desire and Belief verbs. Recall that our model incorporates an attentional mechanism whereby when it encounters an SC utterance during the initial stages, it has some difficulty in encoding the mental event due to a competition arising from the action within the SC. Since Belief verbs more frequently appear with the SC syntax, they are more likely to be mis-interpreted at early stages, giving rise to a delay in their acquisition. In contrast, Desire verbs frequently occur with non-SC syntax. These non-SC usages of Desire verbs are correctly recorded even at the early learning stages. These results point to the importance of looking at both SC and non-SC usages of MSVs. To replicate observations from language learners, our model requires to produce a delay in the association of Desire verbs to their common properties regardless of their usages in non-SC syntax, which the model can record as associating Desire verbs to Desire meaning.

Looking more closely at Figure 5.5(b), we can see that early on our model interprets a Belief test frame mostly as having a Perception meaning, and only sometimes as having a Communication meaning. Interpretation of Beliefs as Perception or Communication initially is consistent with the hypotheses of de Villiers and Pyers (2002) and Israel (2008). However, if we look at the distributions of these three verb classes (Belief, Perception, and Communication) in our data (Figure 5.6), we cannot explain why a Perception interpretation is more likely than Communication: Compared to Perception verbs, Communication verbs seem to have a distribution closer to that of Belief verbs. Clearly, the constructions do not fully capture the interaction among the different verb
classes. In addition, we saw that our model did not show a similar behaviour to children in that it learned Desire verbs too quickly. We attribute this limitation to the fact that the constructions do not capture the interaction between Desire verbs and the other semantic classes (e.g., Action) that could only be captured through generalizations over the full range of syntactic behaviour of all verbs. We now turn to the same analysis using the model’s learned verb classes.

Analysis based on the Learned Verb Classes

Figure 5.7(a) presents the event type likelihoods given the Desire-inf test frame, according to the verb classes. Unlike in the model when using only constructions, here we observe a delay in the association of Desire verbs to their mental meaning, as observed in children (Bartsch and Wellman, 1995). The replication of this trend is enabled by capturing the association of Desire verbs to Communication and Action verbs over the use of transitive constructions as well as infinitival-SC: The similarity of the overall syntactic distribution of Desire verbs to Communication and Action verbs can be seen in Figure 5.6.

Figure 5.7(b) presents the event type likelihoods for the Belief-fin test frame, which show a delayed acquisition compared to Desire verbs (cf. Figure 5.7(a)). We replicate

---

3Note that the class-based likelihoods are shown on a different scale than those based on the constructions. We have fewer verb class clusters, and thus we cannot directly compare the likelihood values. Instead we focus on the trends in the relative likelihoods across the 5 event types.
Chapter 5. An Incremental Model of Learning Verb Classes

this trend when using verb classes, because our incrementally-learned verb classes actually capture the higher similarity of the distribution of Belief verbs across the syntactic patterns to other verb classes, compared with the relatively distinctive distribution of Desire verbs (as can be seen in Figure 5.6). Moreover, in contrast to the pattern presented from using only the constructions, initially the model interprets the Belief-fin test frame as either Perception or Communication (with similar likelihoods), which is more in line with what has been suggested in the psycholinguistic literature (de Villiers and Pyers, 2002). Our model’s verb classes capture the similar distribution over syntactic properties of Belief and Communication verbs, unlike the results presented when using only the constructions.

5.4 Discussion

We present a novel computational model that probabilistically learns two levels of abstractions over individual verb usages: constructions that are clusters of similar verb usages, and classes of verbs with similar distributional behaviour across the constructions. Specifically, we extend the model of Alishahi and Stevenson (2008) by incrementally learning token-based verb classes that generalize over the construction knowledge level. In contrast to the models of Parisien and Stevenson and Perfors et al., our model is incremental, and hence enables the analysis of the monotonically developing classes to show the relation to the development of generalization ability in human learners.

We use our computational model of learning verb classes to show the role of a variety of syntactic constructions in learning the meaning of MSVs. While Barak et al. (2013b) focus on the role of one syntactic construction, i.e., SCs, our results here point to the importance of looking at the distribution of MSVs over the full range of syntactic constructions. Our results show an initial high likelihood of interpreting usages of MSVs as having non-mental meaning based on the interaction of MSVs with other verb classes
based on their syntactic distribution. This can serve as an additional research direction where psycholinguistic studies have mostly focused on the cognitive and pragmatic properties of MSVs as a possible cause for the initial production of MSVs in non-mental meaning.

The focus of this work is on the role of the distribution of MSVs over SC and non-SC patterns in their acquisition. We hope to expand this analysis in the future to additional languages that differ in their distributional properties over the syntactic patterns. Moreover, in a preliminary analysis of the formed constructions and verb classes we note that the similarity in semantic properties across verb classes might also play a role in the learning process of MSVs. Notably, we would like to address the possible use of MSVs in non-mental meaning in CDS (Bartsch and Wellman, 1995; Diessel and Tomasello, 2001). We hope to evaluate the role of semantic properties of MSV usages in future work, while carefully assessing the semantic properties of such usages in CDS over time, including possible parental usages of MSVs in non-mental usages.
Chapter 6

Conclusions

Children acquire MSVs much later than other, lower-frequency, words. One factor proposed to contribute to this delay is that children must learn various semantic and syntactic cues that draw attention to the difficult-to-observe mental content of a scene. We have developed a novel computational approach that enables us to explore the role of such cues. Our model replicates several aspects of the developmental trajectory of MSV acquisition. Importantly, this model relies on incremental statistical learning that enables the analysis of the trajectory of acquisition of MSVs. Moreover, our analysis is performed within the context of the complexity of naturalistic CDS, which consists of a variety of semantic classes and syntactic constructions. Importantly, the computational framework allowed us to simultaneously analyze several factors in the acquisition of MSVs (e.g., the developing attention to mental content, the syntactic properties of MSVs). In this thesis, we also present a novel computational extension to a model of construction learning, which incorporates incremental learning of verb classes. This novel model facilitates research of linguistic phenomena beyond the context of the acquisition of MSVs, such as learning verb alternations. In this chapter we review the contributions of each of the experimental analyses performed as part of this thesis, and offer possible research directions to extend it.
Chapter 6. Conclusions

6.1 Summary of Contributions

The acquisition of mental state verbs (MSVs) relies on numerous cognitive and linguistic skills that include the ability to conceptualize mental states, competence with sentential complement (SC) syntax, etc. Psycholinguistic studies of this acquisition process have often focused on just one of the aspects of the required developing skills to acquire MSVs, i.e., the ability to conceptualize mental states (contrast to studies such as (Papafragou et al., 2007) that considered both linguistic and attention difficulties). However, recent studies have shown that children present difficulties in verbally addressing mental states at a stage they already exhibit an ability to conceptualize mental states. In this thesis, I develop a computational model that simulates this acquisition process while assuming the ability to conceptualize mental states. This model expands previous computational models in two ways. First previous computational models of verb learning have mostly focused on concrete verbs that have visually accessible meanings, such as physical actions: go and eat, while here the focus is on MSVs. Moreover, this thesis addresses the representation and analysis of the SC syntax that is commonly used with MSVs, unlike the previously studied physical verbs that often occur with high frequency syntactic constructions with a single verb/clause, e.g., transitive and intransitive constructions.

In this thesis, we analyze the acquisition of MSVs using a computational model that provides a complementary analytical tool to the many psycholinguistic studies of this process. Our model enables the analysis of MSV acquisition using an algorithm that learns the probabilistic association of verbs and relevant semantic and syntactic properties (see Chapter 3). Notably, we integrate the analysis of linguistic properties of the input along with cognitive and developmental requirements in attending to mental content. The difficulty in attending to the mental meaning in a typical language learning scene is represented in our model by incorporating a probabilistic mechanism of attention that determines how a given scene is semantically interpreted. We show that this developing attention mechanism enables the replication of the delayed association of MSVs to their
typical linguistic properties (i.e., mental semantics and SC syntax). Our results extend
the psycholinguistic experiments by showing the importance of correctly recording the
used syntactic structure despite the difficulty in identifying the main event of the scene,
that is, recording the use of an SC with an MSV.

Importantly, the results of this study imply that psycholinguistic tasks that aim to
evaluate the ability of a child to conceptualize mental states may need to independently
validate the child’s proficiency with the SC syntax. The ability to record mental states is
typically assessed based on the participant’s reply to a verbal utterance that uses an MSV
in an SC construction. Our results indicate that this assessment method incorporates
several layers of difficulty that should be simplified by independently measuring the ability
of the participant to observe and predict actions within the context of a scene and the
ability to verbally describe them. These results are inline with recent psycholinguistic
studies that show evidence for developing an ability to conceptualize mental states when
modifying the experimental settings to rely on eye gaze etc. (e.g., (Rubio-Fernández and
Geurts, 2012)).

In Chapter 4, we focus on the role of the distributional properties of CDS in facilitating
or hindering the acquisition of verbs, and in particular the acquisition of MSVs (Barak
et al., 2013a,b). We do so while considering two sub-groups of MSVs, i.e., Belief and
Desire verbs. Using our model, we study a key assumption of usage-based theories:
that the acquisition of a construction relies heavily on the existence of a high-frequency
exemplar verb that accounts for a large proportion of usages of that construction in
the input (Casenhiser and Goldberg, 2005). Importantly, unlike the psycholinguistic
experiments that focus on the learning of an artificial novel construction using novel verbs,
here we examine the acquisition of the English SC construction from naturalistic input.
Specifically, we examine the effect of overall verb frequency, as well as the frequency
with which a verb and its semantic class appear with an SC. Our results provide new
insights into exemplar-based learning in the context of naturalistic input with multiple
semantic classes, and a diverse set of constructions for the verbs. Our results imply that psycholinguistic experiments that aim to analyze exemplar-based learning should control such additional properties of the experimental data beyond the overall verb frequency. Moreover, studies such as that of Miller (2006) discuss language intervention that aims to improve the ability of children to attend to mental states by exposing them to related linguistic properties. Our results indicate the distributional properties that may facilitate better association of MSVs to the appropriate linguistic properties in such language intervention tasks.

We then further our investigation on the role of the distributional properties of the input more specifically in the acquisition of the two sub-classes of MSVs, namely Desire and Belief verbs. The acquisition of Belief verbs lags behind the acquisition of Desire verbs in children. Some psycholinguistic theories attribute this lag to conceptual differences between the two classes, while others suggest that syntactic differences are responsible. Through computational experiments, we show that a probabilistic verb learning model exhibits the pattern of acquisition, even though there is no difference in the model in the difficulty of the semantic or syntactic properties of Belief vs. Desire verbs. Our results point to the distributional properties of various verb classes as a potentially important, and heretofore unexplored, factor in the observed developmental lag of Belief verbs.

This difference in the distribution of all verb classes over the two syntactic structures (finite vs. infinitival-SC) leads to an early and prolonged association of the finite-SC to non-Belief meaning in our model. These results may imply a state of competition between the association of Belief meaning to other possible meanings used with this syntax, e.g., Communication meaning. Previous studies have suggested that verb classes, such as Communication verbs, assist the acquisition of Belief verbs by offering a more visually accessible interpretation of mental scenes as a partial initial meaning, e.g., interpreting thinking as expression of perspective or point-of-view on an event, similar to saying (de Villiers, 2005). Indeed, our results show that Belief verbs are initially associated to
Chapter 6. Conclusions

non-Belief meaning based on the association of finite SC to other semantic interpretations. Although this initial association may be interpreted as delaying the association of Belief verbs to the correct and full meaning (i.e., Belief meaning), it may be essential to access the mental meaning by initially associating MSVs with a more salient perceptual meaning of the scene.

The acquisition of MSVs has been extensively studied in respect to their common occurrence with SC syntax both in the psycholinguistic literature and in our own work here. However, MSVs also occur in a variety of other syntactic structures. Our study in Chapter 5 aims to expand the experimental work presented in the previous chapters by considering the less-explored property of the usage of MSVs with non-SC syntax. Moreover, other verb classes frequently occur with SCs, e.g., Communication and Perception verbs. The similarity in distribution of the various verb classes over syntactic patterns may affect the acquisition of the meaning of MSVs by association. We analyze the association of mental verbs to their meaning over a variety of syntactic patterns using a novel model of verb class learning. Our results in this chapter point to an important role of the full syntactic preferences of MSVs on top of their occurrences with SCs.

The contribution of this chapter goes beyond the experimental analysis it presents in its development of a novel computational model to learn verb classes. We develop an incremental Bayesian model that simultaneously and incrementally learns argument structure constructions and verb classes given naturalistic language input. Our model enables the analysis of monotonically growing constructions and classes to shed light on the trajectory of acquisition processes. Moreover, we present preliminary analysis that supports the contribution of our model as a computational tool that overcomes several of the shortcomings of previous models of verb class learning.
6.2 Future Extensions

The work presented in this thesis can be extended in several directions that range over various levels of implementation and conceptual complexity. The rest of this section describes some of these directions roughly ordered by the expected level of complexity of their implementation:

**Automatic annotation of CDS input.** The analysis performed in this thesis holds the advantage of being carried in the context of naturalistic input generated following the methodology of Alishahi and Stevenson (2008). This method enables the representation of rich syntactic and semantic properties that captures a variety of syntactic constructions, semantic properties of the event and the participants, etc. However, the input is generated to reflect the distributional properties of this limited selection of verbs given a limited number of manually annotated utterances per each verb. Therefore, this method limits the number of verbs that can be included in the data given the tedious annotation required for each verb. An accurate representation of the use of a construction in CDS may rely on a higher number of verbs. For example, the prepositional dative construction studied in Section 5.2 is frequently used with low frequency verbs. Therefore, many low frequency verbs are required to accumulate substantial number of usages of this construction. Another shortcoming of our input generation method is reflected in the accuracy of the representation for each verb. The size of the sample annotated for each verb bounds the range of syntactic and semantic properties observed for this verb. For example, *think* has 100% frequency with the finite-SC in our input generation lexicon based on the sampled data for this verb, even though *think* can occur in transitive construction as in “I think so”.

In order to address these shortcomings of our input generation method, we can use

---

1We estimate the number of non-SC usages of *think* in the data around 5% based on automatic annotation of the syntactic patterns. Given the high frequency of *think* in the input (e.g., 13829 occurrences), the sample of 100 usages per verb is unlikely to capture the infrequent constructions the verb occurs with.
automatically annotated input. Syntactic annotation methods can be used to generate such input while taking under consideration the consequent higher error rate compared with manual annotation, especially for multi-verb utterances including the SC syntax. The semantic annotation presents a challenge of higher complexity compared with the syntactic annotation, since semantic properties often rely on the specific make-up of the scene. A preliminary input can be constrained to syntactic properties, while a limited range of semantic properties can be extracted from VerbNet to describe the semantic roles of the event and arguments in respect to the frame.

**Rich syntactic representation.** The analysis of MSV acquisition required us to develop a syntactic representation that would address the unique properties of SC syntax (e.g., inclusion of two verbs, and independent feature for prepositions/predicates, i.e., use of “NP V that SC” vs. “NP V who SC”, etc.). We represent the syntactic properties of a verb usage using the **syntactic pattern** feature (e.g., “arg1 verb arg2”) and a range of additional features that varies across the representation approaches described in each chapter (e.g., **argument count**, **verb count**, **complement type**). The variations in the representations aimed to minimize the assumptions we made regarding the language proficiency of the learner, while marking the unique properties that distinguish the various constructions. However, throughout the thesis, the syntactic pattern conveys the core information by providing an overall picture of the number, type, and order of the participating arguments and verbs. Following Alishahi and Stevenson (2008), the syntactic pattern is compared across usages as a whole, i.e., as a boolean value of equality between two patterns.

As part of our research, we relaxed the syntactic representation to some extent by representing the prepositions and predicates independently of the pattern. Based on this representation, the syntactic pattern “arg1 _verb_ PP SC” can be derived from both “She thinks that daddy is sleeping” and “I know why he is sad”. However, the model still cannot capture partial similarity across patterns or produce partial patterns based on
an observed verb usage similar to a human response. For example, the model would not produce “daddy is sleeping” based on an observed usage of “She thinks that daddy is sleeping” by omitting the head clause. To simulate partial comparison and production, the model would need to consider partial similarity across patterns as it does for partial similarity of semantic properties. Notably, the syntactic pattern should be compared while taking into consideration the order of the arguments, and the dependencies across the parse tree to guarantee proper alignment of the patterns. That is, a partial syntactic representation would require a development of a precise similarity measure of syntactic patterns.

**Knowledge driven change in attention.** Chapter 3 describes our simulation of the developing ability to attend to the mental content. Our model simulates this change as a function of time, i.e., the likelihood of correctly attending to mental content increases over the course of training. Future research should aim to achieve the gradual change in attention to mental events as an outcome of the learning process. This outcome could be modeled as a function of the accumulated data in the clusters. For example, the attention to concepts can be estimated based on their entrenchment in the data while taking advantage of the lower frequency of mental events in CDS compared with physical actions. The model can be more likely to interpret a scene as expressing a previously encountered physical action rather than a novel mental state given the familiarity of the model with the more frequent physical actions rather than the infrequent mental states. The interpretation of a scene as mental should gain likelihood as the model encounters more occurrences of MSV usages and learns the distinguishing properties of usages of such verbs.

**Individual differences in language acquisition.** This thesis focused on the acquisition of MSVs by children with normal language development. However, the acquisition of MSVs has been studied specifically with children with developmental disorders, such as autism, given its role in referring to the mental state of others. MSVs may be
harder to master for such children given the complexity of their informational and linguistic requirements, regardless of the conceptualization factor. Therefore, some studies offer intervention programs for children with developmental disorders that focus on the use of the common linguistic properties of MSVs independently of the mental context (e.g., the use of SC) (Miller, 2006). Our model can be used to analyze hypothetical processing difficulties that are relevant to such disorders and measure the ability to replicate the observed patterns in the acquisition of MSVs. Moreover, language intervention can be simulated by training our model on input that mimics the hypothetical beneficial distributional properties. This method can shed light on the required distributional properties of the input that result in a closer learning pattern of normally developing children.

**Performative usages.** A vast amount of psycholinguistic research has questioned the role of performative usage in the acquisition of MSVs. In such usages MSVs do not indicate mental content, but rather direct the conversation, e.g., *think* can be used as a politeness marker, as in “I think we should go”, without depicting a genuine mental state (Diessel and Tomasello, 2001) (see Section 2.3.1 for full discussion of performative usages). Our model can be used to explore the role of this stage in the acquisition of MSVs: whether it simply reflects the distribution of performative usages of MSVs in CSD, or whether it precedes full acquisition of MSVs in full mental meaning as a result of some difficulty in verbally referring to the mental meaning of the verb.

This research direction relies on further analysis of CDS input as well as modifications to the model. First, the analysis of the performative stage requires annotation of performative usages in CDS that specifies their frequency for each verb independently. Second, the semantic and syntactic properties of the model should represent such usages of MSVs. Although many performative usages use non-SC syntax, e.g., “I guess”, Bartsch and Wellman (1995) and Diessel and Tomasello (2001) note that even SC usages of MSVs may be considered as performatives (see 2.3.1 on page 15 for examples). Therefore, the semantic properties should distinguish non-mental usages of MSVs from
the mental usages independently from the syntactic properties. Third, the model would need to simulate the processing demands of the various usages of MSVs. Israel (2008) associates the performative stage with the possible limitations in the memory and resulting processing abilities of children. According to this theory, performative usages hold lower processing demands since the head clause conveys simpler meaning when it does not refer to mental states. Modeling such processing limitations and development may go beyond the study of the performative stage to more general questions regarding the interaction of the cognitive and linguistic development over the course of language acquisition.

**Referential uncertainty.** Alishahi and Stevenson (2008) applied their model to utterances with a single verb (e.g., *go* in “daddy *went* to the kitchen”) associated with a set of corresponding event primitives (e.g., \{ *physical, act, move* \}). However, in a typical language learning scene many actions may occur simultaneously. For example, the child may observe daddy *going* to the kitchen, while at the same time daddy was *holding* a cup and *talking* on the phone.

In this thesis, we expanded the level of complexity represented in the input compared with Alishahi and Stevenson (2008) by including utterances with an SC, i.e., with two verb and two sets of event primitives. These utterances required us to address the various interpretations a language learner might give to the usage given the use of two verbs. Yet, we did not consider the referential uncertainty that is likely to arise from a complex language learning environment. While utterances with SC represent usages of two verbs that ”stand in a semantic relation to each other, typical language learning scene requires consideration of several independent sets of semantic properties in respect to each verb usage. In future work, our model can be used to analyze this aspect of learning verb meaning by representing the variability in the semantic interpretation each verb usage may get given the variety of possible associated meanings among the actions and states occurring in the scene.

**Verb class model.** Chapter 5 presents a study of learning verb alternations us-
ing our novel extension to the model to learn verb classes. The analysis presented in this chapter only provides replication of the experimental setup used by previous studies (Perfors et al., 2010; Parisien and Stevenson, 2010). Following previous studies, the input to the model in Section 5.2 relies only on the syntactic properties of the data. However, psycholinguistic studies have also considered the role of the semantic properties of the verbs in the formation of verb alternation knowledge (Ambridge et al., 2012). Moreover, the model is currently evaluated by measuring its generalization ability in respect to the dative alternation. The flexibility of our model and our input representation enables a much broader analysis of verb alternation learning in respect to other common alternations and other aspects of generalizations across verb classes. Our model can be used to study high level generalizations over constructions in the presence of rich semantic properties, verb alternations, etc.
Bibliography


Chris L Baker, Rebecca R Saxe, and Joshua B Tenenbaum. 2011. Bayesian theory of


