Hierarchical Bayesian models of verb learning in children

by

Christopher Parisien

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

Copyright © 2011 by Christopher Parisien
Abstract

Hierarchical Bayesian models of verb learning in children

Christopher Parisien
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2011

The productivity of language lies in the ability to generalize linguistic knowledge to new situations. To understand how children can learn to use language in novel, productive ways, we must investigate how children can find the right abstractions over their input, and how these abstractions can actually guide generalization. In this thesis, I present a series of hierarchical Bayesian models that provide an explicit computational account of how children can acquire and generalize highly abstract knowledge of the verb lexicon from the language around them. By applying the models to large, naturalistic corpora of child-directed speech, I show that these models capture key behaviours in child language development. These models offer the power to investigate developmental phenomena with a degree of breadth and realism unavailable in existing computational accounts of verb learning.

By most accounts, children rely on strong regularities between form and meaning to help them acquire abstract verb knowledge. Using a token-level clustering model, I show that by attending to simple syntactic features of potential verb arguments in the input, children can acquire abstract representations of verb argument structure that can reasonably distinguish the senses of a highly polysemous verb.

I develop a novel hierarchical model that acquires probabilistic representations of verb argument structure, while also acquiring classes of verbs with similar overall patterns of usage. In a simulation of verb learning within a broad, naturalistic context, I show how
this abstract, probabilistic knowledge of alternations can be generalized to new verbs to support learning.

I augment this verb class model to acquire associations between form and meaning in verb argument structure, and to generalize this knowledge appropriately via the syntactic and semantic aspects of verb alternations. The model captures children’s ability to use the alternation pattern of a novel verb to infer aspects of the verb’s meaning, and to use the meaning of a novel verb to predict the range of syntactic forms in which the verb may participate. These simulations also provide new predictions of children’s linguistic development, emphasizing the value of this model as a useful framework to investigate verb learning in a complex linguistic environment.
Dedication

For Pa, a man who could always surprise you
with what you never thought he knew.
Acknowledgements

A thesis does not grow in a vacuum. I owe a debt of gratitude to many people who have helped me greatly over the years. To my supervisor, Suzanne Stevenson, I am deeply grateful for your guidance. You taught me to cope with the stress and uncertainty of research, and you knew when I needed a steady hand. You taught me how to be critical of my work, but also to be confident and proud of what I do. To my advisory committee, Graeme Hirst, Rich Zemel, and Gerald Penn, thank you for your stimulating comments, your insightful critique, and for letting me wander through answers to questions that were never asked. To my external examiner, Dan Roth, thank you for taking the time to visit our group, and for ideas I wish I had thought of five years ago.

I thank my collaborators, Afsaneh Fazly, Daphna Heller, Martha Palmer, Claire Bonial, Susan Brown, and Jena Hwang. You gave me an opportunity to be a part of something truly new and interesting, and working with you has been a source of great joy. I am also grateful to have been a part of an inspiring group of students. To Afra Alishahi, Vivian Tsang, Paul Cook, Aida Nematzadeh, Libby Barak, Frank Rudzicz, Uli Germann, Saif Mohammad, Tim Fowler, Xiaodan Zhu, Mike Demko, Faye Baron, Jackie Cheung, Abdel-Rahman Mohamed, Julian Brooke, Siavash Kazemian, Varada Kolhatkar, Tong Wang, and everyone else, thank you. I have long believed that to do the best work you can, surround yourself with the best people. You have given me this.

To my parents, my brother, and my extended family, both old and new, thank you for your love, support and encouragement. You taught me the excitement in knowledge, and you delighted in letting me share that with you, whatever my fascination might be.

To Sarah, the love of my life, thank you for being my strength. When I come home with a great set of results, you’re there to celebrate with me. When I have a bad day, you’re there to help me. You give me perspective, and you teach me that above all, family matters.
## Contents

1 Introduction .................................................. 1

2 Verb learning in children and in computational models .................. 8
   2.1 Theories of argument structure acquisition .......................... 9
       2.1.1 Bootstrapping theories ........................................ 10
       2.1.2 Usage-based learning ....................................... 14
       2.1.3 Weak abstract representations ............................ 17
   2.2 Computational models ........................................... 18
       2.2.1 Learning verb argument structure ............................. 19
       2.2.2 Verb classes ................................................ 23
       2.2.3 Other Bayesian models of language acquisition ............ 25

3 Argument structure and the acquisition of verb polysemy ................. 28
   3.1 Related work .................................................. 31
   3.2 Verb usage clustering .......................................... 32
       3.2.1 Verb features ............................................... 33
       3.2.2 Dirichlet process mixture model ............................ 34
       3.2.3 Parameter estimation ..................................... 37
   3.3 Experimental set-up ............................................ 38
   3.4 Experiments and analysis of results ................................ 40
   3.5 Richer syntactic features ....................................... 44
3.6 Summary of contributions ........................................ 46

4 Joint model of argument structure and alternation classes 48
4.1 Related work ..................................................... 51
4.2 Description of the model ....................................... 52
  4.2.1 Verb features .................................................. 53
  4.2.2 Model 1: Argument structure constructions ............... 54
  4.2.3 Model 2: Alternation classes .............................. 57
  4.2.4 Parameter estimation ........................................ 59
4.3 Experimental set-up ........................................... 60
4.4 Experiments and analysis of results ............................ 61
  4.4.1 Experiment 1. Verb argument preferences .................. 61
  4.4.2 Experiment 2. Novel verb generalization .................. 63
4.5 Discussion of the model ....................................... 66
  4.5.1 Relation to existing topic models ......................... 66
  4.5.2 Sequential data and exchangeability ...................... 68
  4.5.3 Noise .......................................................... 71
  4.5.4 Relation to Alishahi and Stevenson (2008) ............... 71
4.6 Summary of contributions ..................................... 72

5 Generalizing between form and meaning 75
5.1 Model description .............................................. 78
  5.1.1 Representation of Verb Usages .............................. 78
  5.1.2 The hierarchical model of verb knowledge ................. 79
5.2 Experimental set-up ........................................... 81
5.3 Experiments and analysis of results .......................... 83
  5.3.1 Experiment 1. From alternations to verb meaning ........ 83
  5.3.2 Experiment 2. From verb meaning to alternations .......... 89
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>Summary of contributions</td>
<td>92</td>
</tr>
<tr>
<td>6</td>
<td>Conclusions</td>
<td>94</td>
</tr>
<tr>
<td>6.1</td>
<td>Summary of contributions</td>
<td>94</td>
</tr>
<tr>
<td>6.2</td>
<td>Future directions</td>
<td>96</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Other forms of semantic generalization</td>
<td>97</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Acquisition of verb arguments</td>
<td>97</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Other problems in language acquisition</td>
<td>98</td>
</tr>
<tr>
<td>6.2.4</td>
<td>Probabilistic VerbNet</td>
<td>101</td>
</tr>
<tr>
<td>A</td>
<td>Dirichlet process mixture model</td>
<td>103</td>
</tr>
<tr>
<td>B</td>
<td>Verb class model inference</td>
<td>106</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>113</td>
</tr>
</tbody>
</table>
# List of Tables

3.1 Coarse-grained senses of *get*. .................................................. 30

3.2 Slot features. ................................................................. 35

3.3 Precision (P), recall (R) and F-measure (F) for each sense of *get*. .......................... 41

3.4 F-measures for varied amounts of data, simulating order of acquisition. ................. 43

3.5 Precision (P), recall (R) and F-measure (F) from clustering the data with enriched syntactic features. ........................................... 45

4.1 Slot features. ................................................................. 54

5.1 Training conditions for the novel verb in Experiment 1. ....................................... 85
List of Figures

3.1 Likelihood of grouping usages from each pair of senses. 42
4.1 Hierarchical structure of the models. 55
4.2 Argument preferences for known dative and benefactive verbs. 62
4.3 Generalization of novel dative verbs. 65
5.1 Hierarchical structure of the model. 79
5.2 Using the alternation and animacy of a novel verb to infer meaning. 87
5.3 Using the meaning of a novel verb to constrain its usage. 91
B.1 Plate diagram of the verb class model. 107
Chapter 1

Introduction

One of the most incredible feats that children accomplish is learning to communicate effectively with those around them. Natural language is immensely complex, yet by the age of 5, typical children can hear fully adult utterances and mostly understand them. What is particularly remarkable about language is that it is so deeply productive—even toddlers can easily understand and say things they have never heard before. This productivity in language depends on a large amount of apparent structure organizing a language in coherent ways. Learning how to be productive with language requires two key skills in development. Children must acquire the abstract structures that represent the way their native language works, and they must learn how to generalize that abstract knowledge in appropriate ways. Verbs play a key role in this developmental process.

By representing an event, a verb draws together many different participants in a scene into a cohesive relational concept. The kinds of syntactic arguments it can take are generally constrained by the kind of actors and other participants in that event. It comes as no surprise, then, that a substantial body of work in computational linguistics, cognitive linguistics and psycholinguistics is built on the notion that the syntactic behaviour of a verb is very closely tied to the verb’s meaning. This relationship between form and meaning leads to regular structure in the verb lexicon, as classes of verbs that mean sim-
ilar things tend to be used in similar ways. This has implications for the computational acquisition of a verb lexicon, as well as for our understanding of how the human mind learns about and represents verbs. By attending to regular patterns in the syntactic and semantic behaviour of verbs in usage, a learner can acquire useful linguistic abstractions. Moreover, by generalizing this abstract knowledge appropriately, it is possible to support further learning and capture aspects of the productivity that is so central to natural language.

Computational models of verb learning in children can benefit the broader developmental and computational linguistics research communities in two parallel ways. Firstly, they provide a test bed for developmental theories of language. By implementing computational systems based on these theories and fleshing out the necessary details, researchers can assess the feasibility of various mechanisms in child verb learning, in naturalistic contexts. Since a child has access to very little “labelled” training data, if any, such models must be unsupervised. By doing so, these models provide insight into representations and inference mechanisms to manage the uncertainty and complexity in real-world language use, and can be used to further develop linguistic theories. Secondly, these unsupervised models can be used to develop linguistic resources for other computational methods, without the need for costly annotated resources. By automatically building a detailed verb lexicon from real-world corpus data, we can offer a viable alternative to the laborious manually-constructed resources typically in use today.

To support these two parallel goals, the research in this thesis draws heavily from a theoretical perspective known as usage-based language acquisition (e.g., Tomasello, 2000, 2003; Langacker, 2000; Goldberg, 2006). In contrast to the view that certain aspects of language are an innate part of the human organism (e.g., Chomsky, 1981; Pinker, 1989), thus simplifying the learning problem, usage-based theorists argue that abstract linguistic structures can be learned solely by attending to patterns of language use in the child’s input. The results in this thesis provide further support for the claim
that children’s linguistic input is rich enough, and that general inference mechanisms are powerful enough, for them to acquire and generalize highly abstract linguistic knowledge from the language to which they are naturally exposed. In Chapter 2, I provide more detail regarding this theoretical perspective and competing views.

In this thesis, I focus on two distinct levels of abstraction in verb knowledge. The first level, that of argument structure, is concerned with the syntactic and semantic relationships among a verb and its arguments. When a verb is used with a specific set of arguments, what does that indicate about the meaning of the verb in that context? What is an appropriate representational form for this knowledge in order to acquire it from usage? The second level of abstraction centres on the notion of alternations in verb argument structure: verbs show different patterns in how they can express their semantic arguments in syntactic forms, and the range of different argument structures that a verb can take bears a strong relationship to the meaning of the verb itself. As I will discuss below, this relationship between verb alternation structure and verb meaning is a rich and potentially very useful source of information with which to develop a verb lexicon.

The main hypothesis of this work is that the probabilistic representation of verb classes provides a strong explanation for how children can acquire and generalize abstract knowledge of verb alternations from usage. In a broad context of acquiring knowledge about verb usage, jointly modelling the acquisition of verb argument structure and verb classes permits us to investigate and explain specific generalization behaviours in children, for which no computational account has previously existed. Specifically, the use of nonparametric Bayesian models supports the acquisition of robust representations of abstract verb knowledge in a broad context of naturalistic child-directed speech. Such probabilistic methods can exploit the gradience and flexibility inherent in language to provide robust reasoning about complex, real-world data. The family of Bayesian models I consider in this thesis provides an ideal framework for combining probabilistic reasoning with a complex, hierarchically organized lexicon.
In this dissertation, I explore this hypothesis by building up progressively more powerful Bayesian models that acquire and generalize abstract verb knowledge from naturalistic corpora of child-directed speech. I present these as competency models showing what can be acquired and how it can guide general inference, but I do not make claims regarding the specific processing mechanisms involved. In Chapter 3, we\textsuperscript{1} employ a Dirichlet process mixture model (DPMM; Neal, 2000) to acquire probabilistic representations of argument structure. To explore the notion of systematic relationships between argument structure and verb meaning, we determine how well the model can use the syntactic aspects of argument structure to distinguish the various senses of the polysemous verb get, comparing the model’s behaviour with that of young children.

Chapter 4 investigates ways to acquire a much broader range of verb knowledge. Competent language speakers make very subtle judgments about the ways in which verbs can be used with a range of arguments, adjuncts and syntactic expressions. The different ways in which verbs can express their semantic arguments in syntactic forms, known as diathesis alternations, are often accompanied by subtle changes in meaning. In a broad study of diathesis alternations of English verbs, Levin (1993) showed that regularities in verb alternation patterns correspond strongly with regularities in meaning. Classes of verbs that can be used in similar ways tend to mean similar things. Several computational linguists have exploited this relationship in automatic learning of verb information (e.g., Korhonen et al., 2003; Schulte im Walde, 2008; Vlachos et al., 2009), but very little work has been done on computational models of how children could acquire such complex information.

One commonly studied behaviour in child language involves the English dative alternation. Verbs which allow the dative alternation can occur in either a prepositional dative form (1.1a) or in a double-object dative (1.1b):

\textsuperscript{1}Throughout this thesis, I employ the pronoun I to refer to the thesis writer, and we to refer to the authors of previously published collaborative work in which the thesis writer was involved.
(1.1)  a. I gave a toy to my dog.

    b. I gave my dog a toy.

Many English verbs of transfer that take both a Theme and a Recipient object occur in the dative alternation, including such verbs as *give, take, send, carry, throw,* and *sell.* However, certain subclasses of verbs are unlikely to demonstrate the alternation, including Latinate verbs (*I donated the museum the money*) and verbs of manner of speaking (*John shouted Mary the answer*) (Gropen et al., 1989; Levin, 1993).

Developmental studies have suggested that 3-year-old children are sensitive to the abstract structures involved in the dative alternation (Thothathiri & Snedeker, 2008), and that such children can even generalize knowledge of the alternation to novel verbs of transfer (Conwell & Demuth, 2007). When taught a novel verb in one form of the dative alternation, for example, by hearing an experimenter say, *Look! I gorped you the truck!*, paired with a new kind of transfer action, 3-year-old children will often describe the action using the *alternating* form of the expression—*I gorped the truck to you.*

We aim to show how the abstract information relevant for this generalization can be acquired from observations of how verbs occur with individual arguments in the input. We present two distinct models based on the Hierarchical Dirichlet Process (HDP; Teh et al., 2006). By attending to the syntactic aspects of verb argument use, as in the previous chapter, the first model acquires probabilistic argument structures across a wide range of verbs, generalizing its acquired structures over multiple verbs. We then develop a novel extension to this model that addresses its limitations by also learning classes of similar verbs. By comparing the behaviour of these two models to that of children, we show that usage-based models must include the inference of verb class structure, not simply the inference of individual constructions, in order to account for the acquisition of verb alternations.

In Chapter 5, we pursue a deeper investigation into the relationship between form and meaning in the acquisition of verb argument structure and alternations. As men-
tioned above, verbs that are used in similar ways tend to mean similar things. Another good example of this is the verb *break*, which participates in the causative/inchoative alternation (Levin, 1993):

(1.2)  
   a. Jimmy broke the television.  
   b. The television broke.

Here, the object of the transitive (the Patient, *the television*) becomes the subject of the intransitive. This alternation pattern commonly applies to unaccusative verbs denoting a change of state, like *bend, freeze, or dry*. Now consider the verb *laugh*, which also occurs in both transitive and intransitive forms:

(1.3)  
   a. Janice laughed her glee.  
   b. Janice laughed.

In this case, the subject of the transitive (the Agent, *Janice*) remains the subject of the intransitive, and the object of the transitive is dropped. Moreover, the intransitive form is much more frequent than the transitive. Unlike the change of state verbs listed earlier, this alternation pattern is typical of expression verbs like *cry, snort* and *giggle*.

Children have been shown to be sensitive to these regularities between form and meaning in verb learning. At two years of age, children can use the alternation structure of a novel verb to infer aspects of what that verb means (Naigles, 1996; Scott & Fisher, 2009). By hearing a novel verb used in the pattern of example 1.2, these children can infer that the verb may denote some kind of change-of-state action. Moreover, somewhat older children can use aspects of a novel verb’s meaning to predict the ways it might (or might not) be used (Ambridge et al., 2011). For example, these children can predict that a novel change-of-state verb can acceptably be used in a transitive frame, without ever hearing the verb used in that way.

We augment the verb class model of Chapter 4 with a set of semantic features to more fully account for the regularities between form and meaning that appear to influence verb
learning in children. The model acquires probabilistic associations between syntactic and semantic aspects of verb argument structure. By also acquiring alternation classes of verbs, as in Chapter 4, it is able to perform high-level generalizations of this knowledge to novel verbs, simulating child behaviour. We show that the model can use the syntactic alternation structure of a novel verb to infer aspects of its meaning, and use the meaning of a novel verb to predict its range of acceptable syntactic forms. By making key predictions of child behaviour, we show that the model is a useful framework to investigate the interaction of complex factors in verb learning.
Chapter 2

Verb learning in children and in computational models

As discussed in the previous chapter, our cognitive representations of verbs appear to depend on very strong abstract regularities in the lexicon. These abstractions are central to the productivity of language. They allow us to generalize our linguistic knowledge to new situations, in order to understand and produce language in novel ways. The emergence of generalizations in language development signals important changes in children’s representation of linguistic knowledge. To understand the developmental process behind these changes, we must investigate both how children can find the right abstractions over their input, and also how these abstractions can actually guide the process of generalization.

In this chapter, I review two complementary perspectives on research into these questions. In Section 2.1, I discuss the predominant theories of how children manage to develop the verb knowledge necessary to become mature speakers of their language. These theories are, necessarily, abstractions of a complex developmental process operating in a complex environment. Computational models have the potential to manage a great deal of this complexity, offering a precise way to test and refine developmental theories. In Section 2.2, I review a series of computational models that researchers have used to
investigate the acquisition of verb argument structure, verb classes, and other linguistic phenomena. By fleshing out the details in developmental theories, researchers can employ such models to test the feasibility of specific theories and can offer predictions of human behaviour to stimulate further research. Ideally, we should aim to sustain a dialogue between researchers conducting child experiments and computational modellers, continually refining our understanding of development from both perspectives.

### 2.1 Theories of argument structure acquisition

Children face an interesting problem when learning verb argument structure in their native language: how to rule out unseen constructions as ungrammatical, rather than just unseen. For example, many verbs permit the dative alternation, occurring in both double-object and prepositional dative forms. From observing a few dative verbs, it might be reasonable for a child to expect that all verbs that occur in one form can occur in the other. Of course, this is untrue in general — *confess*, for example, can occur in one construction, but not the other. Despite never being explicitly taught that *confess me your sins* is ungrammatical, competent English speakers have no trouble avoiding the form. How would a child learn that this double-object form is ungrammatical, rather than just not yet seen? This concern, known as Baker’s Paradox (Baker, 1979), highlights the problem of negative evidence, that children receive little to no explicit evidence that certain grammatical constructions are not permitted in the language. It is part of a broader argument in language acquisition known as the Poverty of the Stimulus (see, e.g., Chomsky, 1981). The essence of the argument is that the input children receive contains too little information to learn a mature, final-state grammar for a child’s native language. The solution, as argued by Chomsky and others, is that some aspects of language are cross-linguistically universal, and that variation among languages is dictated by innately defined choices. These universal principles and properties of language could be an innate
part of the human biological organism. Thus, given some innate constraining principles of what a language could be, children may be born with enough knowledge to help them “break into” the learning problem. Here, I review how nativist theories of language propose that children learn about verb argument structure. I then discuss the dissenting view that innate linguistic knowledge is unnecessary for language development, and that solely by using general cognitive skills, children can acquire mature language capabilities from their input alone.

2.1.1 Bootstrapping theories

According to nativist views of language acquisition, innate linguistic principles may take multiple different forms. Children may be endowed with knowledge of basic syntactic categories such as nouns and verbs, core semantic roles such as Agent and Theme, and expectations about linking rules connecting semantic roles to syntactic positions. For example, children may have an expectation that the Agent of an action is likely to be the subject of the verb. Given these innate building blocks, children “bootstrap” themselves to develop a mature, abstract system of verb argument structure. They may use semantic rules to infer regularities in syntax, or they may use syntax to guide the learning of semantics. These two directions of inference are not necessarily mutually exclusive, and this section outlines the motivations and criticisms of each aspect.

Semantic bootstrapping

Under a semantic bootstrapping account of the acquisition of verb argument structure, children first focus on the semantics of the event denoted by a verb, then by aligning the semantic structure with the verb’s syntactic arguments, come to learn about regularities in argument structure (Pinker, 1989). For example, a child might observe that in a breaking event, there is an Agent that causes something to break and a Theme that breaks. When the child hears John broke the car, the innate linking rules would specify
that the Agent (John) is the subject and the Theme (the car) is the object. Children can learn exceptions to the innate biases, for example by recognizing that in the alternating form The car broke, the Theme is the subject.

A child would develop general rules for alternations by observing that some verbs, like break, can occur in more than one lexicosemantic structure, then forming a broad rule to derive one structure from another. In this case, the child would infer that when a causative event involving a thing is expressed as a transitive (John broke the car), the embedded dynamic event (where something is happening to the car) can be expressed as an intransitive, and vice-versa. These broad-range rules are expected to be highly productive, and account for many of the innovative usages of verbs by children and adults. More specifically, Pinker argues that children’s overgeneralization errors are often one-shot innovations resulting from these broad-range rules, such as when children overextend the causative alternation to say things like Adam disappeared the ball. Under these rules, specific alternations are extended on very general grounds, for example, to any kind of dynamic event.

As children use these highly generative broad-range rules, they also gradually form more narrow-range semantic classes of alternating verbs, which allow them to overcome rampant overgeneralization errors. A child might observe that a particular verb alternates, then would extend this property to other semantically similar verbs. By learning that give participates in the dative alternation, she might extend the alternation to other verbs of transfer, or from open, she might extend the causative transitivity alternation to other change-of-state verbs. With sufficient experience, these narrow-range semantic classes become more refined and accurate, and over time they serve to preempt the more general broad-range rules. By recognizing that disappear belongs to a specific narrow-range semantic class that preempts the broad-range rule above, a mature child can retreat from the overgeneralization disappear the ball. In spontaneous speech and elicited production studies, Pinker and his colleagues found that children’s generalizations and
overgeneralizations of alternations largely respect the constraints given by semantic verb classes (Pinker, 1989; Gropen et al., 1989). For example, the passive voice tends to extend to novel verbs of action, but not to novel verbs of experience.

Over the years, several points of criticism have emerged regarding the semantic bootstrapping hypothesis. One important point is that in of the earliest utterances of English-speaking children, the syntactic subjects are not Agents (e.g., I like it, Pete hurt by car; Bowerman, 1990). These are early violations of the supposed innate linking rules, strong biases that should take time to overcome. Furthermore, Goldberg and her colleagues argue that innate linking rules are unnecessary for observing semantic bootstrapping effects (Goldberg et al., 2004). Instead, for any given construction, children may use the most frequent verb associated with that construction as a “pathbreaking” verb, associating the semantics of that verb with the construction. For example, the verb give is typically the most frequent verb used in a double-object construction in child-directed speech. This construction can itself become associated with the meaning of give, such that verbs with meaning similar to give (such as show) could also acceptably be used in a double-object form.

**Syntactic bootstrapping**

A shortcoming of semantic bootstrapping in explaining acquisition of verb argument structure is that semantics is not always reliable enough to use as a cue to learn syntax. For example, some verbs, like chase and flee, refer to nearly identical situations, differing on only one aspect which is difficult to distinguish just from the event (Gleitman, 1990). In the same event, a fox may chase a rabbit, while the rabbit flees the fox. If chase and flee are used to interchangeably describe the same events, then children cannot learn the difference between them simply by observing their use over many situations. The solution, argued by Gleitman, Fisher, and others, is that children use the syntax of an utterance to help guide their interpretation of the event (Gleitman, 1990; Fisher, 1994,
2000). Under this approach, children attend to the number of a verb’s arguments and their syntactic arrangement, then use this information to frame the semantic structure of the event, helping them infer the meaning of the verb. When hearing *The fox chased the rabbit*, since *fox* is in the subject position, children could infer that the event is most likely “about the fox,” thus framing the event in a different perspective than if the speaker said *The rabbit fled the fox*. Here, syntactic argument position influences meaning. The number of arguments also matters: from hearing the utterance *The dog gorp the cat the ball*, children may infer that *gorp* involves transfer, as in *give* or *throw*.

The syntactic bootstrapping hypothesis has support from several verb learning studies with children. Naigles (1990) used the preferential looking paradigm to show how 24-month-old children can use argument structure to infer the meaning of a verb. Children were shown a video in which an actor dressed as a rabbit pushes another actor, dressed as a duck, into a squatting position. At the same time, both characters circle their arms. As the children watched the scene, one group heard the commentary *The rabbit is gorp the duck* (a transitive) while the other group heard *The rabbit and the duck are gorp* (an intransitive). Later, children were shown two videos side by side with the actions separated, *i.e.*, one with the pushing action and the other with the arm-circling, and told to *Find gorp!* The children who heard the transitive looked at the pushing action, while those who heard the intransitive looked at the arm-circling. Since the transitive generally implies that some object is causally affected by an agent (which is not the case for the intransitive), the pushing action provides the best fit in that case. This result suggests that the children used the argument frame of the novel verb to direct their attention to relevant components of the event. In a later experiment, Naigles (1996) showed that children can also use the multiple syntactic frames in an alternation to infer the meaning of a verb. Using a similar paradigm as in the above study, Naigles showed that children who heard a novel verb used in a causative alternation (*The duck is sebbing the frog/The frog is sebbing*) were likely to infer that the verb involved physical
causation. When the novel verb was used in an omitted object alternation (*The duck is sebbing the frog/The duck is sebbing*), children were more likely to infer that the verb involved physical contact without causation. Naigles argues that in these demonstrations, children use the syntax of the utterance to guide verb meaning. In later work, Scott and Fisher (2009) showed that children can make the same inference without even watching the candidate events unfold. The authors presented novel verbs to 28-month-old children in the appropriate alternations, without an accompanying video. By only hearing the range of syntactic frames in which the verbs occur, the children could still infer the right semantics of the verbs.

### 2.1.2 Usage-based learning

The bootstrapping effects observed in children, as described in the previous section, do not necessarily require innate linguistic knowledge. In contrast with nativist theories of development, usage-based theories argue that abstract linguistic structures can be learned solely from patterns of language use in the child’s input. Language-specific principles are unnecessary for this process; rather, children’s linguistic input is rich enough, and their general cognitive mechanisms are powerful enough, to learn argument structure generalizations from the input using abilities like pattern finding, category learning and problem solving (Tomasello, 2000, 2003).

A number of empirical studies have suggested that children’s early linguistic knowledge is strongly item-based (see Tomasello, 2000, for a review). That is, children’s early language is completely structured around specific words and phrases, not reflecting general syntactic categories or schemas. In particular, this item-specific knowledge appears to be largely organized around individual verbs. Tomasello and others observed that in the speech of two-year-old children, productivity in verb use is very uneven. One verb may be used in only a single construction (*i.e.*, verb-argument configuration, as in *Cut the paper, cut some string, cut my finger*), whereas a very similar verb could be used in
a wide variety of constructions (e.g., *I draw on the man, Draw it by Santa Claus*). A mature representation of argument structure would recognize that both of these verbs permit many different constructions, but young children appear to recognize only specific constructions permitted for specific verbs. Where studies have compared these patterns with adult data, the patterns of caregiver usage closely matched child productivity (e.g., Campbell & Tomasello, 2001). Tomasello and others argue that at this stage, children rely completely on verb-specific constructions with open nominal slots. Other than a category of nominals, children at this stage possess no other linguistic or syntactic abstractions. There is no evidence for innate linguistic categories or rules. Children would not employ such verb-general representations as Agent and Theme, but would instead apply verb-specific roles such as “hitter” and “hittee”.

Children younger than three years are reluctant to use verbs in ways other than how those verbs have already been heard. In an experiment with children aged 2;0 and 2;6 (years;months), Tomasello and Brooks (1998) taught children two novel verbs, one as a transitive (*Ernie is tamming the car*) and another as an intransitive (with a Theme subject: *The ball is meeking*). Even with much encouragement, two-year-olds could not use the verbs in any other way than how they had been heard, and the older children were only slightly more productive. Under the weird word order paradigm, Abbot-Smith et al. (2001) described events with familiar and novel verbs using an incorrect argument order (e.g., *Ernie Bert pushing*), then asked children to describe similar events. Children younger than 3;6 tend to correct the experimenter’s word order for familiar verbs, but with novel verbs, they either use the weird argument structure or avoid the verb altogether. This suggests that, contrary to older children and adults, they do not possess a strong abstract representation of the argument structure constructions to regularize their use of the novel verbs. When children do generalize constructions in novel ways, Abbot-Smith et al. (2004) show that the strength of this generalization is influenced by the input frequency of the constructions.
Langacker (2000) presents a grammar framework to support this kind of usage-based learning. Under this framework, abstract linguistic units are gradually built up by observing regularities in the input. Langacker envisions a bottom-up process where the same cognitive processes apply from the most basic to the most general units, unifying the learning mechanisms and providing a continuum of representation through all domains of language structure. In Langacker’s terms, a unit is a stand-alone concept, some kind of psychological event that has occurred enough times that it can be processed as a “pre-packaged” entity — among other things, these may be events or entities in the world (such as [DOG], a semantic representation of the animal), or phonological items (e.g., [dog], the word). Units can be composed into more complex units, perhaps giving semantic associations ([DOG]/[dog]), or syntactic constructions (e.g., a ditransitive pattern [[V][NP][NP]]). New combinations can be created on-the-fly, and if they occur often enough, they will be entrenched into new units, formalizing the role of input frequency.

In this usage-based cognitive grammar, every abstract representation is grounded in the patterns of the input.

While Langacker’s usage-based grammar certainly allows for associations between form and meaning, in Goldberg’s (2006) framework, form-meaning mappings are essential to the representation of language. Goldberg defines a construction as any linguistic pattern where “some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist.” (Goldberg, 2006, p. 5). Other predicable patterns may be stored as constructions if they occur with sufficient frequency. Constructions occur at all levels of linguistic abstraction, including words, idiomatic constructions like believe ⟨one’s⟩ ears, or argument structure constructions like the ditransitive. Goldberg claims that by combining constructions, we can impart new meanings to words that would not reasonably be stored as such in the mental lexicon. Before hearing something like She whooshed into my life, we would not normally expect whoosh to be used as a verb. In this case, the argument frame suggests metaphorical
motion, as in *She flew into my life*. The argument structure transfers the semantics of motion to the word *whoosh*, so we understand the sentence as intended.

Goldberg’s theory offers an explanation for children’s recovery from overgeneralization errors. By a process of *statistical preemption* (Goldberg, 1995), a child can avoid an overgeneralization by recognizing that a word is more likely to occur in a competing pattern. Basic preemption occurs with Pinker’s broad-range and narrow-range rules — as a child learns the more specific narrow-range rules, these rules preempt the broad rules, reducing the overgeneralizations that come with them. Here, *statistical* preemption accounts for the influence of input frequency. For example, causation can often be conveyed with the transitive construction, but *He laughed her* and *He giggled her* are overgeneralizations. With sufficient input, these errors can be preempted by the periphrastic causative: *He made her laugh, he made her giggle*. Since the former occurs more frequently than the latter, this theory predicts that *He giggled her* would be a more common error.

### 2.1.3 Weak abstract representations

Shortly after the presentation of Tomasello’s (2000, 2003) usage-based theory, a debate emerged with Fisher (2002a) regarding the nature of children’s early representations of language. Under a strong version of the usage-based theory, Fisher argues, children should not apply any generalizations of verb argument structure until they have formed the appropriate abstract representations, which apparently occurs around age three. However, Fisher points out that while there is a strong tendency for younger children to be conservative in their verb use, a non-trivial number of two-year-olds in the above-mentioned studies *do* in fact generalize novel verbs to new sentence frames, or correct the word orders of novel verbs. This contradicts the strong version of the theory. Furthermore, recent studies have demonstrated children’s use of argument structure and alternations in ways that show young children can generalize abstract structure, just not in a fully mature way. For example, Conwell and Demuth (2007) showed that when
taught a novel verb in the double-object form of the dative (*He pilked Petey the cup*), three-year-olds can produce the verb in the alternating prepositional form (*He pilked the cup to Petey*). The children were far less robust making the opposite generalization, suggesting that the abstract alternation pattern was not yet fully formed. Using a syntactic priming paradigm, Thothathiri and Snedeker (2008) showed that while young children’s abstract representations of dative constructions may not be strong enough to influence production, the constructions can still influence comprehension.

To explain these observations, Tomasello and Abbot-Smith (2002) suggest that one need not take the strong form of the usage-based theory as assumed by Fisher (2002a). Rather, children may possess weak representations that allow some forms of generalization but not others. As children are exposed to more input, these representations would strengthen over time. Current research in the field now aims to characterize the nature of these weak abstract representations (*e.g.*, Abbot-Smith et al., 2008), although some researchers do claim that evidence of the linking rules central to bootstrapping theories appears before children begin generalizing argument structure (Fernandes et al., 2006).

### 2.2 Computational models

Computational models provide an opportunity to test and refine theories of language acquisition. The availability of large corpora of child-directed speech such as CHILDES (MacWhinney, 2000) has allowed modellers to examine how much linguistic structure is learnable from the input children receive. In this section, I review several computational models of the acquisition of verb argument structure, verb classes, and other related aspects of language.

Since the current state of the art in computational models of language acquisition focuses on usage-based learning, I will not cover models in the Principles and Parameters (P & P) framework in depth. Important models in this field include Clark’s (1992) use
of a genetic algorithm for parameter setting and Gibson and Wexler’s (1994) Triggering algorithm. The unifying theme in much of this work is that P & P intends to restrict the possibilities of language variation, thus reducing the acquisition problem to a search over a finite hypothesis space. Considering the child’s need for a universal mechanism to process an utterance for its input to the learning algorithm, Fodor (1998a, 1998b) argues for the existence of a universal parser as part of the innate language organ.

2.2.1 Learning verb argument structure

An early and highly influential model of word learning was given by Siskind (1996), who presented the first implemented model of cross-situational inference, that is, the acquisition of form-meaning mappings by tracking how lexical forms tend to co-occur with different objects and events across many situations (Quine, 1960; Pinker, 1989). Siskind views the word learning problem as one of finding an alignment between the lexemes of an utterance and a set of logical forms hypothesized for a given scene. A single pairing of a scene and an utterance will produce a very large number of possible mappings, but by intersecting these sets of hypotheses over many situations, the model can restrict those mappings to those that are logically consistent. Siskind’s model does not make use of any syntactic properties of language to guide its conceptual alignments. Moreover, its rule-based approach is limited in its applicability to naturalistic data. Once a word is “learned” in this model, its meaning cannot be revised, so it is extremely sensitive to noisy data. The model is tested on artificially generated input rather than naturalistic data.

Several models have sought to demonstrate the interactions between syntax and semantics described in bootstrapping theories of acquisition, but without innate linguistic knowledge. Desai (2002) gives a connectionist model that learns to associate lexical and syntactic forms with semantic features. The model takes artificial input generated from a simple grammar, covering three basic argument structures: an intransitive (a boy is
jumping), an intransitive with two subjects (a boy and a girl are jumping), and a simple transitive (a boy is pushing a girl). The model uses backpropagation to associate the lexical features of the input with scene descriptions in the output, including the objects in the scene, causality, and other features of the event. By learning to associate the transitive frame with causality, the model is able to infer that a novel transitive verb is likely to be causative, thus showing syntactic bootstrapping behaviour. The authors also claim that since knowledge of the meanings of the nouns in an utterance aids learning of a novel syntactic frame, the model also shows limited semantic bootstrapping. However, the model requires a very large number of training instances, and it is unclear how the model would perform if given more complex input.

Niyogi (2002) offers a Bayesian account of bootstrapping effects. Using a highly simplified representation of some syntactic and semantic features of verb usages, the model predicts a verb’s class from observations of its use. The class defines acceptable semantic and syntactic patterns for the group of verbs, allowing these features to be generalized to novel verbs as appropriate. The model can make reasonable inferences about novel verbs from a small number of observations, but it relies on large amounts of hand-coded verb knowledge. The modeller must specify the structure of the hypothesis space (i.e., the contents of the known verb classes), the conditional probabilities of features given each of those classes, and the prior distribution over classes. The model does not specify how this knowledge may be learned; thus it is not an acquisition model.

Jones et al. (2000) use a network model to demonstrate how children may acquire syntactic frames linked to specific lexical items, as suggested by Tomasello (2003). The model takes child-directed speech from the CHILDES database as input (MacWhinney, 2000), then incrementally adds links and nodes to the network to represent common patterns in the input. For example, given utterances Want to eat and Want to play, the model would create a structure representing the frame Want to X, with extensions eat and play. As predicted by the verb-island hypothesis (Tomasello, 2003), the model
acquires a large number of frames linked to specific verbs (e.g., get + N, want + N), and relatively few frames centred around common nouns (e.g., baby + V). The model does not make use of any semantic information.

In English, subject-verb-object ordering is so prevalent that verb arguments can often be identified by word order alone. This is not a universal property of language. In Italian, for example, sentence constituents have relatively free order, and overt subjects may be omitted (i.e., Italian is a pro-drop language). While in English, word order is the most effective cue for identifying subjects and objects in child development, Italian children tend to rely on noun-verb agreement and noun animacy (Bates et al., 1984). Using a Maximum Entropy model, Dell’Orletta et al. (2005) demonstrate how these features can be used to acquire subject-object identification in Italian from an annotated corpus. Their results demonstrate a prediction of some usage-based learning theories, that high-frequency light verbs (such as, in English, go, give, and want) serve as “pathbreaking” verbs to generalize from local verb-specific patterns to the language as a whole.

As suggested by Goldberg (2006), argument structure constructions pairing form and meaning can explain how syntax and semantics interact in a usage-based model of language. Some computational models have sought to explain how these constructions might be learned and used. One example of this is the work of Chang (2004), who uses a unification-based grammar formalism to show how form-meaning schemas can be built up from more basic relations in the input. Given an ontology of schemas for actions, locations, objects, and so on, as well as a set of simple lexical constructions linking words to concepts, Chang’s algorithm processes child-directed speech to build various abstract constructions. A THROW-TRANSITIVE construction, for example, might connect lexical arguments of throw to verb-specific semantic roles such as a “thrower” and a “throwee.” As suggested by the verb-island hypothesis, these early constructions tend to be tied to specific verbs. The model performs limited generalization, for example by abstracting the THROW-BLOCK and THROW-BALL constructions into a general THROW-OBJECT
Alishahi and Stevenson (2008) show, using a general Bayesian clustering framework, how argument structure constructions can be learned in a way that reflects the generalization behaviours of young children. The authors represent each verb usage in the input with a frame containing that token’s argument structure and a simple representation of the scene, including object roles and event semantics like causation and motion. Using an incremental clustering algorithm based on general human category learning (Anderson, 1991), their model clusters similar verb usages together to learn common form-meaning associations. The model simulates typical overgeneralization errors in children (e.g., an incorrect transitive Adam fall ball), and demonstrates how statistical preemption (Goldberg, 2006) can allow children to overcome those errors. By associating semantic features with argument structure, the authors show syntactic bootstrapping behaviours (as in Naigles, 1990), where a novel verb blick, used in a transitive form (The bunny is blicking the duck), is likely to be causative.

BabySRL (Connor et al., 2008) is a semantic role labelling model that examines the structure-mapping aspects of the syntactic bootstrapping hypothesis. The model employs a simple linear classifier to classify semantic roles by the positions of noun arguments. The authors argue that simple features, denoting the number of noun arguments and their positions relative to each other, can show the generalization behaviours of 21-month-old children. These children tend to assume that the first argument is an Agent and the second is a Theme, even in the intransitive phrase The dog and the cat are pushing. By transitioning to a slightly more complex feature which denotes the argument position relative to the verb, the model overcomes this error, as do children. While this model requires supervised training data, Alishahi and Stevenson (2010) give an extension of their 2008 model to simulate the unsupervised learning of general semantic roles.
2.2.2 Verb classes

Several computational systems have demonstrated the value in using argument structure information to learn about verb semantics, although many of these are not models of human learning. Schulte im Walde (2008) clusters verb types using a variety of deep and shallow syntactic relations, including grammatical relations from a full parser as well as the part-of-speech classes of words in a context window. By comparing the clustering results against GermaNet and FrameNet classes and human associations, she finds that simple co-occurrence features can perform as well as, and in some cases better than, deep syntactic relations. Korhonen et al. (2003) use subcategorization frame distributions to cluster verbs with a combination of nearest-neighbour and information-theoretic approaches. An advantage of their approach is the use of soft-clustering methods, allowing polysemous verbs to belong to more than one verb class. Vlachos et al. (2009) also use subcategorization frame distributions to cluster verbs, this time applying a Dirichlet Process Mixture Model to gain the flexibility of learning an unspecified number of verb clusters. Lapata and Brew (2004) make use of subcategorization frames to improve verb sense disambiguation. Given a partially parsed corpus, for each combination of a verb token and its syntactic frame, their model generates a probability distribution over Levin’s verb classes (Levin, 1993). This use of argument structure to infer verb class makes a valuable prior distribution for disambiguation systems.

Joanis et al. (2008) examine a verb classification task using syntactically shallow slot features — subject, direct and indirect object, for example. The authors also add simple semantic features to their representations, addressing other important aspects of verb alternation behaviour: the overlap of arguments across syntactic slots; tense, voice, and aspect; and the animacy of individual arguments. Classification results show that shallow syntactic features of the arguments are the most useful for this task, while semantic features are indeed helpful in some cases. In particular, features approximating semantic roles are necessary to distinguish certain alternation classes, a point made clear
by Merlo and Stevenson (2001). For example, unaccusative verbs, denoting changes of state, permit both an intransitive and a transitive form (e.g., *The butter melted/The cook melted the butter*). In the intransitive form, the subject is the Theme, while in the (causative) transitive form, the subject is the Agent and the object is the Theme. Object-drop verbs also permit intransitive and transitive forms, yet in both cases, the subject is the Agent (e.g., *The boy played/The boy played soccer*). The object, taking the Theme role, is optional. Such verb classes cannot be distinguished by subcategorization frames alone, yet Merlo and Stevenson show, using a corpus of *Wall Street Journal* text, that the addition of simple features such as the animacy of the subject can successfully discriminate these classes. Scott and Fisher (2009) replicate this experiment using a corpus of child-directed speech, showing that the same distributional cues are available to children. Scott and Fisher also demonstrate the importance of using distributional information to distinguish verb classes: 28-month-old children were able to infer part of the meaning of a novel causal or object-drop verb simply by listening to how the verb alternates in sentences, that is, without watching the event.

Perfors et al. (2010) use a hierarchical Bayesian model (HBM) to explain how children may use distributional statistics of the input to determine which verbs may or may not participate in the dative alternation and how knowledge of these classes may be generalized to novel verbs. To distinguish alternating from non-alternating dative verbs, their model estimates the variability of the constructions with which different verbs occur. For example, since the dative verb *give* occurs in both double-object and prepositional dative constructions, it would be represented in a different class from the Latinate verb *donate*, which does not occur in the double-object form. The feature-variability model on which their work is based is not specific to language, but rather has been used to simulate a variety of cognitive phenomena, including the shape bias and other human category-learning behaviours (Kemp et al., 2007). However, the model of Perfors et al. (2010) assumes that the individual constructions participating in the dative alternation
have already been learned. Furthermore, the authors limit their model to only consider two possible constructions (the prepositional and double-object dative), and only the verbs that participate in those constructions.

These models of verb clustering and classification typically compare verbs on the basis of the argument structure constructions with which they occur — verbs that occur with similar constructions are considered to belong to the same class. Gries and Stefanowitsch (2010) take a slightly different approach by considering the strength of association between verbs and constructions. This approach, which they dub *collostructional analysis*, examines the likelihood with which verbs occur with particular constructions. If a verb $v$ occurs with construction $C$ (e.g., the ditransitive) significantly more often than verbs in the corpus at large, then we say that $v$ is a significantly attracted *collexeme* of $C$. For any given construction, Gries and Stefanowitsch (2010) cluster verbs by the strength of this association, and find that the resulting clusters often correspond with common semantic features such as communication or physical force.

### 2.2.3 Other Bayesian models of language acquisition

Since this thesis is focused on the use of Bayesian models of language acquisition, we consider several recent models that have been developed in other areas of language acquisition that may give useful insight to the task of learning verb alternation classes. Goldwater et al. (2009) demonstrated a model of word segmentation in infants that learns to cluster sequences of phonemes into words based on statistical regularities in the input. The model uses a nonparametric clustering technique that is similar to the approaches in some of the above models, but includes the temporal ordering of the input as a key factor in the structure of the model.

Yu (2006) considers the syntactic-semantic interactions of bootstrapping theories of acquisition in the broader context of word learning. Yu argues that learning syntactic categories such as nouns and verbs can help general word learning by grouping and
highlighting words that are likely to have the same type of referent. By integrating these syntactic cues into a Bayesian model, Yu shows how syntactic-semantic mappings can improve the accuracy of word learning. Xu and Tenenbaum (2007) present a different Bayesian framework for word learning that accounts for the hierarchical nature of words for object categories. If, for example, a word is used to label a goldfish, a Dalmatian and a parrot, the model will expect the word to apply to the category of animals as a whole. However, if a word is used to label a goldfish, a salmon and a trout, the model will infer that the word most likely applies to the more specific category of fish. The model appropriately accounts for humans’ uncertainty in learning category labels in these tasks.

Other models address the problem of referential uncertainty in word learning. Fazly et al. (2010) develop a probabilistic model of cross-situational word learning based on alignment models in machine translation. As with the intent of Siskind’s (1996) approach, their model incrementally processes scene-utterance pairs in the input to learn associations between words and semantic elements. Using a large and noisy corpus of child-directed speech, the model simulates several important phenomena in word learning, including the vocabulary burst and fast mapping behaviour. An extension to the model (Alishahi & Fazly, 2010) examines the influence of syntactic categories on the word learning task. Frank, Goodman, and Tenenbaum (2009) show a similar model that incorporates social cues into cross-situational word learning. Given a set of possible referents in a scene, a speaker will typically refer to only a small subset of those elements in an utterance. This model uses cues such as eye gaze to help infer the speaker’s intended referent, improving the likelihood that it will learn the correct mapping between a word and its meaning. Unlike the model of Fazly et al. (2010), this model uses a relatively small vocabulary. Moreover, the model of Frank, Goodman, and Tenenbaum (2009) processes all input as a batch, rather than incrementally, which reduces its cognitive plausibility.

Frank, Goodman, Lai, and Tenenbaum (2009) exploit the Gricean maxim of “being informative” to improve a model of word production and word learning. If a speaker
wishes to identify an object out of a large group of items, she will generally pick out that object by describing the properties that best distinguish it from the rest of the group. Under this model, the speaker's intent is represented by a probability distribution over possible meanings (that is, a distribution over the objects in the domain). The speaker’s chosen words also provide a distribution over meanings—how the possible referents in the domain fit the referring expression. The speaker chooses the words that give the best fit between her intent and what the words convey. The authors argue that by taking this as an expectation of what a speaker will do, such a model can improve word learning by effectively reducing referential uncertainty.
Chapter 3

Argument structure and the acquisition of verb polysemy

Children’s acquisition of the senses of polysemous verbs has become an important target of study in cognitive linguistics and developmental psychology (e.g., Nerlich, Todd, & Clarke, 2003; Theakston, Lieven, Pine, & Rowland, 2002). Some of the highest frequency and earliest learned English verbs, like put, make, get, go, and do, are also among those with the largest number of senses (E. V. Clark, 1996). Children as young as two years of age freely understand and use many of these polysemous verbs, often with little apparent confusion (Theakston et al., 2002; Israel, 2004). Computational models can help to elucidate the kinds of mechanisms capable of distinguishing the senses of massively polysemous verbs from very little input and the linguistic features necessary to achieve this.

Information about verb senses has been said to correlate strongly with verb argument structure. Several computational models have been developed that make use of a verb’s possible arguments to identify semantic structure and similarity to other verbs. Most of these models operate at a coarse-grained semantic level, grouping verb types into general classes of similar verbs (e.g., Merlo & Stevenson, 2001; Joanis et al., 2008; Versley, 2008;
Korhonen et al., 2003; Vlachos et al., 2009). On the other hand, computational models of child language acquisition have found success by clustering word usages (e.g., Jones et al., 2000; Chang, 2004; Alishahi & Stevenson, 2008), that is, individual instances of verbs along with their contexts. In this chapter, I argue that such usage-based models can be used to study children’s acquisition of verb polysemy.

Because a verb’s arguments serve to identify the participants in a scene, drawing them together into a cohesive event, the syntactic structure of those arguments can indeed be a very strong cue to verb meaning. This chapter is based on the work of Parisien and Stevenson (2009). It is a preliminary investigation into the use of a nonparametric Bayesian clustering algorithm to acquire knowledge of verb argument structure, and to draw semantic inferences from that acquired knowledge.

We analyze the English verb *get* as a case study. *Get* is a particularly interesting target since it is highly frequent, highly polysemous, and is one of the first verbs children learn (E. V. Clark, 1996). Table 3.1 outlines the major senses of *get*. We adopt distinctions made in previously published literature. We adopt the first six senses (*obtain* through *cause become*) from the work of Israel (2004), a corpus study of child language. Berez and Gries (2009) conducted a corpus study of adult spoken and written language, using a finer-grained set of sense distinctions. We adopt an additional distinction of theirs, the *must* sense, to account for a large proportion of the usages that do not fit in the other six coarse-grained senses. The relative frequencies shown in this table are estimated from the work of (Berez & Gries, 2009). Other sets of senses may be found in the literature, but this offers a good assessment of the breadth of meaning captured by the verb. Here, we conflate literal and metaphorical senses. For example, the usage *He got an idea* may be analyzed as metaphorical, in contrast with a more concrete usage like *I got a book*. In this work, both cases would fall under the general sense *obtain*. Various infrequent senses are gathered under *other*.

Children tend to learn more-frequent verb senses earlier than less-frequent senses
Chapter 3. Verb polysemy

<table>
<thead>
<tr>
<th>Sense</th>
<th>Freq. (%)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>obtain</td>
<td>52.3</td>
<td>I got a book.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>He got an idea.</td>
</tr>
<tr>
<td>cause obtain</td>
<td>1.3</td>
<td>I got you a book.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>That mistake got me a suspension.</td>
</tr>
<tr>
<td>move</td>
<td>16.5</td>
<td>You should get on that bus.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Get out of here.</td>
</tr>
<tr>
<td>cause move</td>
<td>5.2</td>
<td>It’ll get you to Buffalo.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Get me home.</td>
</tr>
<tr>
<td>become</td>
<td>15.0</td>
<td>Jim got fired.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I’m getting sleepy.</td>
</tr>
<tr>
<td>cause become</td>
<td>2.5</td>
<td>Suzie got Jim fired.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>This gets you excited.</td>
</tr>
<tr>
<td>must</td>
<td>6.3</td>
<td>I’ve got to go home.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>You’ve got to eat soon.</td>
</tr>
<tr>
<td>other</td>
<td>0.8</td>
<td>You get to eat cake!</td>
</tr>
<tr>
<td></td>
<td></td>
<td>They don’t get along very well.</td>
</tr>
</tbody>
</table>

Table 3.1: Coarse-grained senses of get.

(Theakston et al., 2002; Israel, 2004). However, the order of acquisition does not completely follow the frequency ordering, and this shows that something other than the frequencies of these related polysemous senses contributes to the ease of acquisition. This is a challenge for distributional clustering models, where performance is generally improved with greater amounts of data.

In this chapter, we use a hierarchical Bayesian clustering model to group individual usages of the verb get, drawn from a corpus of child-directed speech. We show good
clustering results by using a set of simple, automatically extracted syntactic features. However, we argue that while these features are commonly used in distributional models of verb semantics, they are inadequate to explain order of acquisition behaviour in children.

The remainder of this chapter is organized as follows. Section 3.1 distinguishes this approach from other related models of verb learning. Section 3.2 describes the general modelling framework. Section 3.3 describes the experiment in clustering instances of *get* in child-directed speech, and Section 3.4 compares the model’s behaviour with that of young children’s abilities to distinguish the various senses of the verb. In Section 3.5, I use a corpus of usages of *get* with an enriched syntactic representation in order to explore one possible explanation for discrepancies between the model’s behaviour and that of young children. Section 3.6 summarizes the contributions of this chapter.

### 3.1 Related work

Several recent computational models have demonstrated the value in using argument structure information to learn about verb semantics. Versley (2008) and Schulte im Walde (2008) cluster verb types using various syntactic dependencies such as noun phrases, prepositional phrases, and adverbs. Following on the work of Merlo and Stevenson (2001), Joanis et al. (2008) achieve similar goals using syntactically shallow slot features – subject, direct and indirect object, for example. In each case, the simple argument structure patterns correlate with human judgements of semantic verb classes.

Few approaches explicitly address the problem of multiple senses of a single verb type. The work of Korhonen et al. (2003) uses a soft-clustering method that allows a verb to belong to multiple possible clusters, allowing a degree of polysemy in a verb’s representation. Verbs are clustered by the distribution of their subcategorization frames. If two senses of a verb differ strongly in their subcategorization patterns, the verb will
more likely be distributed across multiple clusters. Vlachos et al. (2009) use similar subcategorization features in their approach, employing a Dirichlet process mixture model (DPMM) as the clustering algorithm to give the flexibility of learning an unspecified number of clusters. In this case a probabilistic soft-clustering is possible, although the authors do not examine this aspect of the model.

Each of these approaches is concerned with type-level clustering of verbs, that is, clustering verbs based on the distributional properties of all the verb’s usages, taken together. The model may recognize that run, skip and walk are similar, and in the case of Korhonen et al. (2003), that run is also similar to flow, as in the river runs east. However, the verb itself is still represented as a single point in distributional space. A token-level method, on the other hand, clusters individual usages of verbs. This way, different senses can occupy distinct representations of the same verb. Evidence from psycholinguistics suggests that such a method may be necessary to fully explain polysemy (Theakston et al., 2002). Very few models address token-level verb clustering. Lapata and Brew (2004) use subcategorization patterns to perform token-level classification (not clustering) of verbs, thereby presupposing a defined set of verb classes. Alishahi and Stevenson (2008) cluster individual verb usages to simulate the acquisition of verb argument structure in children. Their approach of clustering by using basic argument information is similar to ours, although the incremental algorithm is necessarily sensitive to the order of presentation of the input.

### 3.2 Verb usage clustering

In this section, I describe the modelling framework for clustering verb usages into senses. I discuss the feature representations of individual verb usages, then describe the application of a DPMM, a Bayesian clustering framework well suited to models of human category learning.
3.2.1 Verb features

Following from the type-level verb clustering approaches described above, we designed our feature space to capture some of the general argument structure distinctions between verb senses. We primarily use syntactic “slot” features, similar to those used by Joanis et al. (2008), to encode basic argument information about a verb usage. These are not subcategorization frames, but rather a set of individual features that record the presence or absence of syntactic positions – subject, direct and indirect object, for example – that potentially contain verb arguments. In any particular usage, a certain slot may be analyzed as an adjunct rather than a true argument, but we do not attempt to make that distinction. Such slot features are easier to extract than full subcategorization frames, and Joanis et al. (2008) show that in verb classification tasks, subcategorization frames offer no improvement over simple slot features. We do not include any semantic features in this representation because we intend to see to what extent the various polysemous meanings of *get* are distinguishable purely by syntactic argument information.

Table 3.2 presents the 17 features used in our representation. The first 15 are binary features denoting the presence or absence of a syntactic slot.\(^1\) Since our input data is extracted from the CHILDES database of child-directed speech and child language (MacWhinney, 2000), the labels correspond to the grammatical relations used by the CHILDES dependency parser (Sagae et al., 2007). The syntactic argument information we extract from this parsed input is akin to assuming that at this stage of development, a child can make certain distinctions among the syntactic arguments in the input, but may not yet recognize recurring patterns such as transitive or double-object constructions. This parser introduces a substantial amount of noise into the automatically extracted slot features, which we believe is reasonable given the amount of difficulty a child will experience in making consistent distinctions among verb arguments in the input. When

\(^1\)The slot features include CPRED and XPRED, which do not generally occur as arguments to *get*; included usages in the input are likely parsing errors and are considered to be noise.
one of the other relations is a prepositional phrase, the nominal feature PREP denotes the preposition used. The final feature NSLOTS refers to the total number of slots occurring with this verb usage. The following examples show this representation used with utterances from the input:

(3.1) I got mad because you ran away.
\[ \langle \text{SUBJ, PRED, CJCT, PREP = null, NSLOTS = 3 } \rangle \]

(3.2) Maybe Mama will get the apple for you.
\[ \langle \text{SUBJ, OBJ, JCT, PREP = for, NSLOTS = 3 } \rangle \]

### 3.2.2 Dirichlet process mixture model

As stated earlier, the goal of our approach is to learn clusters of verb usages that approximate verb senses. To achieve this, we use a DPMM, a nonparametric Bayesian model that has gained significant attention in the machine learning community (Neal, 2000). A DPMM brings two main advantages over other clustering methods. Firstly, the modeller need not specify in advance the number of clusters necessary to represent the data. This is the “nonparametric” aspect of the model: as part of the learning process, the model itself determines an appropriate number of clusters, dependent on the data. Secondly, the DPMM has been shown to be a good model of human category learning behaviour (Sanborn, Griffiths, & Navarro, 2006). In addition to basic category-learning tasks, DPMMs and related models have successfully been applied to word segmentation (Goldwater et al., 2009) and type-level verb clustering (Vlachos et al., 2009).

A DPMM specifies a probability distribution over possible cluster arrangements of data. In contrast to typical algorithms that seek a single “best” clustering of data points, a DPMM gives a distribution over all possible clusterings. Given the observed verb usage data, we can estimate the parameters of that distribution to find the most likely clusterings.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Feature type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJ</td>
<td>Subject</td>
<td>Mary got a new car.</td>
</tr>
<tr>
<td>CSUBJ</td>
<td>Finite clausal subject</td>
<td>That you left got me upset.</td>
</tr>
<tr>
<td>XSUBJ</td>
<td>Non-finite clausal subject</td>
<td>Watching TV gets boring.</td>
</tr>
<tr>
<td>OBJ</td>
<td>First object</td>
<td>You got a book.</td>
</tr>
<tr>
<td>OBJ2</td>
<td>Second object</td>
<td>You got me a book.</td>
</tr>
<tr>
<td>IOBJ</td>
<td>Prepositional object</td>
<td>Get it to me.</td>
</tr>
<tr>
<td>COMP</td>
<td>Finite clausal complement</td>
<td>I get that you don’t want to be here.</td>
</tr>
<tr>
<td>XCOMP</td>
<td>Non-finite clausal complement</td>
<td>You won’t get me to eat that.</td>
</tr>
<tr>
<td>PRED</td>
<td>Adjectival predicate</td>
<td>Mom got mad.</td>
</tr>
<tr>
<td>CPRED</td>
<td>Finite clausal predicate</td>
<td>The problem is that you’re too tired.</td>
</tr>
<tr>
<td>XPRED</td>
<td>Non-finite clausal predicate</td>
<td>The goal is to win.</td>
</tr>
<tr>
<td>LOC</td>
<td>Locative</td>
<td>Get it here.</td>
</tr>
<tr>
<td>JCT</td>
<td>Adjunct</td>
<td>I got an apple for you.</td>
</tr>
<tr>
<td>CJCT</td>
<td>Finite clausal adjunct</td>
<td>I got scared when you screamed.</td>
</tr>
<tr>
<td>XJCT</td>
<td>Non-finite clausal adjunct</td>
<td>I got tired of eating granola.</td>
</tr>
<tr>
<td>PREP</td>
<td>Preposition (nominal value)</td>
<td>I got it for you.</td>
</tr>
<tr>
<td>NSLOTS</td>
<td>Number of slots used</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Slot features.

We assume that each verb usage $y_i$ belongs to a cluster, and that its features are drawn from a set of multinomial distributions (one per feature). Different clusters are associated with different feature distributions. For each cluster, we obtain the likelihood that a verb usage in this cluster will present each of the different slot features. While we use features of individual slots like OBJ or COMP, each cluster probabilistically represents a subcategorization frame. Thus, one cluster may probabilistically represent
Chapter 3. Verb polysemy

a pattern of features such as SUBJ V OBJ, while another cluster may represent the pattern SUBJ V OBJ COMP. The number of clusters in turn depends on a Dirichlet Process (DP), a stochastic process which gives the model its nonparametric flexibility. The full model is defined as follows:

\[
\begin{align*}
\pi & \sim \text{Stick}(\alpha) \\
z_i & \sim \text{Multinomial}(\pi) \\
\theta_{jz_i} & \sim \text{Dirichlet}(1) \\
y_{ij} & \sim \text{Multinomial}(\theta_{jz_i})
\end{align*}
\]

The \(\sim\) symbol should be read as “is distributed according to”. In the above, \(y_{ij}\) denotes feature \(j\) of verb usage \(i\). \(z_i\) is an indicator variable that identifies the cluster chosen for usage \(i\), and \(\theta_{jz_i}\) are the multinomial parameters for feature \(j\) in the probabilistic pattern represented by the cluster. The Dirichlet is a continuous probability distribution over the parameters of a multinomial. By drawing the \(\theta_{jz_i}\) from this Dirichlet, we place a weak uniform prior on the distributions of individual features. Since \(z_i\) selects \(\theta\) from across the set of clusters (e.g., \(\theta_{j1}\) or \(\theta_{j2}\)), \(\pi\) provides a mixing distribution, defining the prior probability that \(y_{ij}\) belongs to each of the clusters. Finally, \(\alpha\) is a concentration parameter that affects how many clusters we expect to find.

The parameters to this mixing distribution, \(\pi\), are drawn from a stick-breaking process (denoted as Stick) at the core of the model’s nonparametric\(^2\) flexibility. This process gives a prior distribution over the total number of clusters to use and which of those clusters will be more likely overall. It is equivalent to the following stochastic process: assume that all verb usages have been clustered except \(y_i\). Then the prior probability of \(y_i\) being

\(^2\)The term “nonparametric” is often a source of confusion, given that nonparametric models typically have large numbers of parameters. The term is used in contrast to other kinds of mixture models, where the number of mixture components (here, clusters) must be explicitly specified as a fixed parameter by the modeller. In nonparametric models, the number of mixture components is estimated by the model itself. In doing so, the model represents a potentially infinite number of components, and a potentially infinite number of parameters. Accordingly, another appropriate term for this family of models is “infinitely parametric.” The reader may be excused moderate frustration at this state of affairs.
placed in cluster $k$ is given by

$$P(k) = \begin{cases} \frac{n_k}{N-1+\alpha} & \text{if } n_k > 0 \text{ (existing cluster)}, \\ \frac{\alpha}{N-1+\alpha} & \text{otherwise (new cluster)}, \end{cases}$$

(3.3)

where $n_k$ is the number of verb usages in cluster $k$ and $N$ is the total number of usages. Larger values of $\alpha$ make it more likely that overall, more clusters will be used. In all the experiments in this chapter, we set $\alpha = 1$, a moderate setting that compares with similar DPMM applications. This formulation has two interesting properties. Firstly, larger clusters tend to attract more usages. Secondly, as more data is processed, the probability of choosing a new cluster decreases. These properties relate nicely to human category learning: more heavily entrenched categories are more likely to be recognized in the future, and over time, subjects tend to create fewer completely new categories to organize their environment (Anderson, 1991).

The above model, as written, specifies a prior distribution over the complete set of possible parameters to the model (i.e., all possible values for $\theta$, $z$, and $\pi$). To find clusters of verb usages, we update this distribution using the observed data, thus obtaining a posterior distribution over parameters.

### 3.2.3 Parameter estimation

Given the set of verb usage data, we estimate the posterior distributions over the model parameters using Gibbs sampling, a Markov Chain Monte Carlo (MCMC) method (Neal, 2000). Essentially, to estimate a probability distribution, we draw a large number of samples from that distribution. The samples give an approximation of the distribution, and as the number of samples approaches infinity, the approximation becomes exact. With Gibbs sampling, we choose an initial random setting for the model parameters (i.e., the cluster assignments $z$ and the cluster parameters $\theta$), then iteratively adjust these settings according to the observed data.
In this work, we implement the DPMM in OpenBUGS (Lunn et al., 2009), a general framework for MCMC inference in Bayesian models. We do not have direct access to the details of the Gibbs sampling method, though we provide a high-level description here, as follows. In our experiments, we randomly set each $z_i$ to one of a small number of clusters (1, 2, or 3). For each cluster, we set the $\theta$ parameters to random values drawn from a Dirichlet distribution. We iteratively update each $z_i$ and $\theta_{jk}$ individually by drawing it from a posterior distribution conditioned on the data and all the other parameters in the model. In the case of a cluster assignment $z_i$, we do this by sampling a cluster for $y_i$ given assignments for all the other usages, as if $y_i$ were the last usage observed. We may choose a new cluster (as in Equation 3.3), thus potentially changing the total number of clusters. We repeatedly cycle through the model parameters, sampling each $\theta_{jk}$ and each $z_i$ many times. By averaging over a large number of these samples, the posterior approximation converges on the exact solution. In practice, we can achieve a good estimate in a few thousand samples, depending on the complexity of the data and the details of the algorithm.

3.3 Experimental set-up

In our experiments, we use child-directed speech data drawn from the CHILDES database of parent-child interactions (MacWhinney, 2000). We use four longitudinal corpora from the American English component of the database, corresponding to four children: Eve, Naomi, Nina, and Peter. Together, the data cover an age range from 1;2 (years;months) to 4;9. We extract each child-directed utterance of the verb get, then randomly split the utterances into development and test sets (1275 and 1276 utterances respectively), dividing each child’s data equally. The corpora contain part-of-speech tags and syntac-

---

3It is common practice to initialize nonparametric models with a smaller number of clusters than expected, although this has a minimal effect on the results. The different numbers of clusters correspond to different initialization conditions; see Section 3.3.
t
tic dependencies, obtained using an automatic tagger and parser (MacWhinney, 2000; Sagae et al., 2007). As described above, we extract 17 slot features for each usage of *get*. Due to errors in the automatic part-of-speech tagging, parsing and feature extraction, the data contains some noise. Some utterances were dropped when parsing errors prevented extraction of the features, and others contain multiple instances of *get*. The final development set and test set contain 1272 and 1290 usages, respectively. For evaluation purposes, I manually annotated each of the usages with one of eight sense labels, corresponding to the eight senses in Table 3.1. We refer to this labelling as the gold standard.

The OpenBUGS model specification for the DPMM is included in Appendix A. We run five chains with different initial conditions: one chain is initialized with all usages in one cluster, two chains start with two clusters, and two with three clusters. Each chain is randomly initialized as described in the previous section. As with other MCMC methods, as the number of iterations of the sampler increases, the distribution of samples approaches the true posterior distribution. As per standard practice, we run each chain for 60,000 iterations, discarding the first 10,000 samples in order to give the sampler time to reach a reasonable approximation of the posterior (a standard process known as *burn-in*). To reduce correlation in the samples, we keep only every 25th sample, giving 2,000 samples per chain, 10,000 in total.

Each sample contains one clustering of the verb usages. To evaluate the model’s performance, we score each of the 10,000 samples against the gold standard, then average the results over all samples. As a result, the reported scores give a weighted evaluation of the entire distribution of clusterings, not just the single “best” cluster. We evaluate each sample using the cluster F-measure (Larsen & Aone, 1999). Given a single sample, for each sense *s*, we score each cluster *k* as follows. Let *a* be the number of usages in *k* with sense *s*. Let *b* be the total number of usages in the cluster, and let *c* be the total number of usages with sense *s*, over all clusters. Then precision (P), recall (R), and F-measure
(F) are given by:

\[ P = \frac{a}{b}, \quad R = \frac{a}{c}, \quad F = \frac{2PR}{P + R}. \]  

We record P, R, and F for the cluster with the best F-measure for that sense, then report averages over all 10,000 samples.

By using an MCMC mechanism to develop an unsupervised clustering of the input, the inference method seeks to optimize the likelihood of the observed data. In other words, the emergent cluster structure is ideally one that offers a good prediction of held-out data. The model finds clusters by attending only to syntactic aspects of verb argument structure, and we evaluate its success by comparing these clusters with verb sense labels. Our claim in this work is that the emergent clusters (formed on the basis of syntactic argument information) correspond closely enough to distinct senses of the verb that such clusters can be used to distinguish verb senses by the syntactic properties of the verb’s arguments.

### 3.4 Experiments and analysis of results

Table 3.3 presents the results of clustering using the DPMM on the test set usages of get. The model uses on average 5.2 clusters. The more frequent senses, obtain, move, and become, achieve the best performance. The less frequent causative senses show worse clustering behaviour, although the recall scores indicate that the model recognizes some internal similarity among the usages. In these cases, low precision scores suggest that the features of the causative senses are quite similar to those of other senses.

We examine this possibility in Figure 3.1, which shows the likelihood of grouping together verb usages from different senses. Lighter shading corresponds to higher likelihood. We calculate the likelihood of each usage of a given gold standard sense being placed in the same cluster as each other usage of the gold standard senses, taken over all 10,000 samples and averaged over usages within each sense. A perfect clustering would
Table 3.3: Precision (P), recall (R) and F-measure (F) for each sense of get.

give a diagonal matrix. High values along the diagonal roughly translate to high recall, and low values on the off-diagonal indicate high precision. The figure shows that cause obtain, cause move and cause become are frequently grouped together (column 2, rows 2, 4 and 6). One possibility is that the model distinguishes causative meanings from non-causatives based on the larger number of arguments in causative forms, but lacks features that would effectively distinguish the various causative meanings from each other.

A common observation in child language acquisition studies is that the more frequent senses of a verb tend to be the earliest senses children produce (Theakston et al., 2002; Israel, 2004). This role of frequency is unsurprising from a machine learning perspective, since we expect more data to make learning easier. Indeed, we see this effect in the results above: the more frequent senses tend to be easier to learn.

On the other hand, the role of frequency in acquisition is not a hard-and-fast rule. There are notable exceptions that can shed light on distributional semantic methods. Israel (2004) studied the order of acquisition of various senses of get, using the same transcripts as in our own study. Using the same sense categories as ours (excluding our
category *must*), Israel compared the frequencies of senses in child-directed speech with the order in which the children first produce these senses. He notes that, in most cases, what a child hears most frequently, he or she learns quickly. The most common exception is *cause obtain*: despite comprising only 2-3% of the input, children often produce it before far more frequent senses like *become* or *cause move*.

In our own results, *cause obtain* is learned better than expected from frequency alone. Table 3.3 shows that *cause obtain* is learned with a much higher F-measure than the similarly-frequent sense *cause become*, although it is no better than *become* or *cause move*, as Israel’s observations suggest. In order to simulate the learning of verb senses over time in our model, we would ideally apply the model to different sets of age-appropriate child directed speech. However, since the corpora under study here each begin at different ages, we focus on the effect of the *quantity* of input on the acquired senses. We run the
model on different-sized subsets of data, randomly sampled from the test set. Table 3.4 shows F-measures of each of the senses, for 400- and 800-usage subsets as well as the full test set. To replicate Israel’s observations, we should expect to see high scores for cause obtain from small amounts of data, that is, earlier than when the scores improve for more frequent senses like become or cause move. We do not see this effect. Rather, cause obtain shows relatively poor performance for all three dataset sizes. It appears then that while slot features give promising clustering behaviour, in these experiments they do not lend themselves to the kind of order of acquisition effects we observe in child behaviour.

Israel (2004), as well as Gries (2006), have suggested that the acquisition of polysemous verb senses may depend on complex inferential mechanisms on the part of the child. For example, the become sense of get may be a metaphorical extension of the move sense, for which children must observe a metaphorical connection between states and locations. As an explanation for the early acquisition of cause obtain, a child could extend obtain by

<table>
<thead>
<tr>
<th>Sense</th>
<th>N=400</th>
<th>N=800</th>
<th>N=1290</th>
</tr>
</thead>
<tbody>
<tr>
<td>obtain</td>
<td>55.1</td>
<td>53.2</td>
<td>56.9</td>
</tr>
<tr>
<td>cause obtain</td>
<td>22.1</td>
<td>22.4</td>
<td>32.8</td>
</tr>
<tr>
<td>move</td>
<td>34.7</td>
<td>43.9</td>
<td>56.0</td>
</tr>
<tr>
<td>cause move</td>
<td>29.1</td>
<td>35.3</td>
<td>37.1</td>
</tr>
<tr>
<td>become</td>
<td>42.4</td>
<td>49.0</td>
<td>59.0</td>
</tr>
<tr>
<td>cause become</td>
<td>6.7</td>
<td>11.3</td>
<td>11.8</td>
</tr>
<tr>
<td>must</td>
<td>4.1</td>
<td>3.9</td>
<td>5.6</td>
</tr>
<tr>
<td>other</td>
<td>4.4</td>
<td>5.1</td>
<td>7.3</td>
</tr>
<tr>
<td>Number of clusters</td>
<td>2.8</td>
<td>3.6</td>
<td>5.2</td>
</tr>
</tbody>
</table>

Table 3.4: F-measures for varied amounts of data, simulating order of acquisition.
adding a causal agent, a connection which children appear to make quite early (Fisher, 2002b). Our model does not make explicit inferences like these, which may explain why our results do not exhibit the same order of acquisition as in children. However, it may be that the behaviour we see in our model is due to the simplicity of our features, or the noise inherent in using automatically extracted data. Children may attend to some other aspect of the input not captured in our fairly simple feature set, something that helps them to acquire certain senses at an early age from comparatively little input. To investigate this, in the next section we apply our model to a richer set of hand-annotated features drawn from a corpus of adult spoken language.

### 3.5 Richer syntactic features

Berez and Gries (2009) analyzed 600 adult-language instances of *get*, sampled from the British component of the International Corpus of English, ICE-GB. The authors annotated the data with 47 fine-grained senses, which we regroup into the 8 coarse-grained labels of Table 3.3. Each usage has been tagged with 13 features commonly used in verb clustering, drawn from the manual annotations of ICE-GB. These features cover a broad range of phenomena, including verb transitivity, verb form, grammatical relations such as the presence of auxiliary verbs, and clausal features including dependency types and the transitivity of dependent clauses.\(^4\)

By encoding verb arguments and certain semantic relationships among them, transitivity patterns capture more information than subcategorization frames or slot features alone. For example, in the “copula” pattern used in this data, an adjectival or prepositional complement describes a property of the subject, as in, *I got involved with that woman*. This semantic property distinguishes the copula from the syntactically similar intransitive pattern. Since these features are hand-annotated, we expect the data to con-

---

\(^4\)See Berez and Gries (2009) for the full list of features.
Table 3.5: Precision (P), recall (R) and F-measure (F) from clustering the data with enriched syntactic features.

<table>
<thead>
<tr>
<th>Sense</th>
<th>P (%)</th>
<th>R (%)</th>
<th>F (%)</th>
<th>Freq. (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>obtain</td>
<td>68.4</td>
<td>56.7</td>
<td>62.0</td>
<td>314</td>
</tr>
<tr>
<td>cause obtain</td>
<td>2.0</td>
<td>49.9</td>
<td>3.8</td>
<td>8</td>
</tr>
<tr>
<td>move</td>
<td>29.9</td>
<td>44.5</td>
<td>35.7</td>
<td>99</td>
</tr>
<tr>
<td>cause move</td>
<td>13.4</td>
<td>61.3</td>
<td>22.0</td>
<td>31</td>
</tr>
<tr>
<td>become</td>
<td>59.0</td>
<td>23.7</td>
<td>33.8</td>
<td>90</td>
</tr>
<tr>
<td>cause become</td>
<td>4.4</td>
<td>44.1</td>
<td>7.9</td>
<td>15</td>
</tr>
<tr>
<td>must</td>
<td>75.8</td>
<td>99.9</td>
<td>86.2</td>
<td>38</td>
</tr>
<tr>
<td>other</td>
<td>1.2</td>
<td>46.8</td>
<td>2.3</td>
<td>5</td>
</tr>
</tbody>
</table>

tain fewer extraction errors and less noise than our own automatically extracted data. We cluster the verb usages using the DPMM and present the results in Table 3.5, scored as in the above experiments.

Overall, these results show a similar pattern to the experiments on CHILDES data. The more frequent senses, obtain, move, and become, perform reasonably well, while the less frequent causative senses perform poorly. The exception is must, with a remarkably high F-measure of 86.2%. This sense is nearly always used in a form similar to I’ve got to X, with highly consistent auxiliary use, verb form and clausal form, all missing from our simple slot representation.

Even with a richer, manually annotated data set, the clustering results do not exhibit Israel’s key observation that the cause obtain sense can be learned earlier than its frequency might predict. These results suggest that in order to accurately model this pattern in acquisition, we would need either a different type of information, or a different approach to learning. The model’s excellent performance on the must sense shows that
given suitable features, a DPMM is capable of learning an infrequent sense very well. A promising direction for future work would be to determine these appropriate features.

Detailed semantic distinctions may be difficult to capture automatically, particularly given the assumption of a child’s limited linguistic development. One option would be to include argument fillers in addition to syntactic slot features. Such an approach may offer additional developmental plausibility: children may associate verb senses with specific lexical items before they are able to access more general argument types. However, selectional preferences have been shown to be largely ineffective for type-level verb clustering (Joanis et al., 2008), although they may offer some benefit at the token level of our approach. Results from sentence processing experiments show that the semantic category of a subject can bias an adult reader’s interpretation of a verb sense, which in turn predicts argument structure (Hare, Elman, Tabaczynski, & McRae, 2009). It may be possible to incorporate this effect by using a word space model for NP arguments (Baroni, Lenci, & Onnis, 2007), or perhaps a simple animacy feature (Joanis et al., 2008).

### 3.6 Summary of contributions

In this chapter, I described experiments using token-level clustering methods to simulate children’s acquisition of the senses of a polysemous verb. With the English verb *get* as a case study, we used a Bayesian framework to cluster usages of *get* drawn from a corpus of child-directed speech. We showed that simple, automatically extracted syntactic slot features gave reasonably accurate clustering results on the senses of *get*. However, these features were insufficient to account for the order of acquisition of polysemy as observed in children. Children do not show a consistent correlation between frequency and age of acquisition. We showed that even with a more detailed, manually-annotated feature set, clustering results in the model did not reflect child behaviour. This suggests that for a token-level clustering method to accurately model this pattern in child language
acquisition, it would need either a different kind of information or a substantially different learning mechanism.

One other possible explanation for children’s apparent ease in learning certain infrequent verb senses is that children may generalize meaning from other similar verbs. For example, children may recognize that the ditransitive use of *get*, as in *I got you a sandwich*, is similar to that of other benefactive verbs like *buy*, *catch*, or *find*. This class of verbs is systematically used in both causative and non-causative forms, and children may recognize this regularity and use it to their advantage. Children are known to generalize verb argument structure and its associated semantic knowledge across many different verbs, and computational simulations suggest that this is an important factor in children’s ability to learn verbs with such ease (Alishahi & Stevenson, 2008). In the next chapter, I present two computational models of the acquisition of verb knowledge that rely on this sort of generalization to support verb learning. One model acquires representations of verb argument structure that may be generalized across a wide range of verbs, and another model provides a powerful extension to this, also acquiring classes of similar verbs to further support generalization.
Chapter 4

Joint model of argument structure and alternation classes

As discussed in Chapter 2, an important debate in language acquisition concerns the nature of children’s early syntax. On one side of the debate lies a claim that children develop their syntactic knowledge in an item-based manner. This claim of usage-based learning argues that very young children associate verb argument structure with specific lexical items, only gradually abstracting syntactic knowledge after four years of age (e.g., Tomasello, 2003). An alternative claim suggests that young children do indeed possess abstract syntactic representations—i.e., generalizations about the structure of their language that are not necessarily tied to lexical items (e.g., Fisher, 2002a).

Syntactic alternation structure is often considered to be a central phenomenon in this debate. Consider the following example of the English dative alternation:

(4.1) I gave a toy to my dog.

(4.2) I gave my dog a toy.

These sentences mean roughly the same thing, but are expressed in different ways. The first, a prepositional dative, expresses the theme (a toy) as an object and the recipient
(my dog) in a prepositional phrase. The second, a double-object dative, expresses both the theme and recipient as objects and reverses their order. Many English verbs occur in this alternation, such as show, lend, write, and tell. Generally, dative verbs in English tend to convey some notion of transfer. However, certain verbs of transfer do not allow the alternation—verbs such as confess (She confessed her sins to the bishop / *She confessed the bishop her sins) and donate (He donated $1,000 to the museum / *He donated the museum $1,000). In many cases, this pattern may depend on fine-grained distinctions in semantic similarity.

In general, verbs that allow similar alternations in their argument structure often have similar semantics (Levin, 1993). This suggests that alternations reflect much of our cognitive representations of verbs. Furthermore, these regularities appear to influence our language use. In word learning experiments, children as young as three years of age appear to use abstract representations of the dative alternation (Conwell & Demuth, 2007). While this is evidence of abstract syntax at a very young age, it does not necessarily invalidate the usage-based hypothesis, since the abstractions may originate from item-specific representations.

One way to bring these opposing positions together is to demonstrate, using naturalistic data, how to connect a usage-based representation of language with abstract syntactic generalizations. In this chapter, previously published in large part as Parisien and Stevenson (2010), we argue that alternation structure can be acquired and generalized from usage patterns in the input, without a priori expectations of which alternations may or may not be acceptable in the language. We support this claim using a hierarchical Bayesian model (HBM) which is capable of making inferences about verb argument structure at multiple levels of abstraction simultaneously. We show that the information relevant to verb alternations can be acquired from observations of how verbs occur with individual arguments in the input.

From a corpus of child-directed speech, the model acquires a wide variety of argument
structure constructions over hundreds of verbs. Moreover, by forming classes of verbs with similar usage patterns, the model can generalize knowledge of alternation patterns to novel verbs. By acquiring these distinct levels of knowledge simultaneously, the model introduces a degree of representational power that has not previously been available in computational accounts of verb learning. In contrast, earlier models have focused on either the acquisition of the constructions themselves \((e.g.,\) Alishahi & Stevenson, 2008), or the formation of classes over given constructions \((e.g.,\) Perfors et al., 2010). The integration in our model of these two important aspects of verb learning has implications for current theories of language acquisition, by showing how abstract syntactic knowledge can be acquired and generalized from usage-level input.

By acquiring probabilistic representations of this high-level verb knowledge from the input, our model offers a valuable perspective on the notion of “weak” abstract representations, suggested by Tomasello (2003) and by Fisher (2002a). As discussed in Chapter 2, some two-year-old children do, in fact, appear to generalize some form of abstract knowledge of argument structure to novel verbs, albeit in a more constrained manner than mature speakers. Some children may generalize novel verbs to new sentence frames, or correct the word orders of novel verbs (Fisher, 2002a). As shown by Conwell and Demuth (2007), three-year-old children can typically generalize a novel dative verb from the double-object form to the prepositional form, but are far less robust making the opposite generalization. These observations suggest that early abstract representations are somehow weaker than those of mature speakers. By representing knowledge of verb argument structure and alternations probabilistically, this kind of constrained generalization is reflected in the degree of uncertainty about an abstract representation given the input.

Usage-based Bayesian models (including this model and others, such as that of Alishahi and Stevenson (2008)), offer an explanation of how such weak abstract representations can arise from the input and can develop into mature linguistic knowledge.

The remainder of this chapter is organized as follows. Section 4.1 explains how this
model relates to similar Bayesian models of verb learning. In Section 4.2, we describe two distinct models for the acquisition of verb knowledge, one that acquires probabilistic representations of verb argument structure, and a second model that adds a level of abstraction to represent classes of similar verbs. Sections 4.3 and 4.4 describe experiments to demonstrate how the two models acquire verb argument structure from the input and show the role of abstract verb classes in generalizing verb alternation structure. Section 4.5 discusses the relationship between the joint model and the family of probabilistic topic models on which it is based, including the implications of certain modelling assumptions we have made. Finally, Section 4.6 outlines the contributions of this chapter.

4.1 Related work

Previous computational approaches to language acquisition have used HBMs to represent the abstract structure of verb use. Alishahi and Stevenson (2008) used an incremental Bayesian model to cluster individual verb usages (or tokens), simulating the acquisition of verb argument structure constructions. Using naturalistic input, the authors showed how a probabilistic representation of constructions can explain children’s recovery from overgeneralization errors. In another Bayesian model of verb learning, Perfors et al. (2010) cluster verb types by comparing the variability of constructions for each of the verbs. The model can distinguish alternating from non-alternating dative verbs and can make appropriate generalizations when learning novel verbs.

Both of the above models show realistic patterns of generalization, but they operate at complementary levels of abstraction. The model of Alishahi and Stevenson learns argument structure constructions but does not capture the alternation patterns of verbs, while that of Perfors et al. learns alternation classes assuming that the individual constructions participating in the alternation have already been learned. Furthermore, Perfors et al. limit their model to only consider two possible constructions (the prepositional and
double-object dative), and only the verbs that participate in those constructions.

In this chapter, we address both levels of abstraction of the above models. We cluster individual verb usages to learn argument structure constuctions and their patterns of use across many verbs, and we also cluster verb types to learn alternation behaviour, generalizing that behaviour to novel verbs. Moreover, we use representative corpora of child-directed speech to model the acquisition of verb alternation behaviour in the context of many constructions, many verbs, and many alternations.

Vlachos et al. (2009) used a Dirichlet process mixture model to cluster verb types by their subcategorization preferences, but did not address learning the argument structures themselves. Other work has modelled different aspects of the dative alternation, such as how discourse features affect the expression of dative constructions (de Marneffe et al., 2011), yet did not consider how these preferences are learned.

### 4.2 Description of the model

In this section, we discuss the feature representation of a verb usage and develop two contrasting models to show how alternation classes contribute to generalization in verb learning. Model 1 is an adaptation of an existing probabilistic topic model, the Hierarchical Dirichlet Process (HDP; Teh et al., 2006), to the problem of learning verb argument structure. Model 2, a novel extension to the HDP, addresses the limitations of Model 1 by learning verb alternation classes, allowing regularities in construction use to be transferred to novel verbs.

While we noted at the beginning of this chapter that verbs that permit the same alternations often have similar meaning, here we focus on the syntactic aspects of alternation structure. That is, we first investigate the kinds of inferences about verb alternations that children could make by attending only to syntactic aspects of verb usage. In the next chapter, we will investigate semantic inference in verb learning by using an augmented
Chapter 4. Joint model

form of the input features described below.

4.2.1 Verb features

As in Chapter 3, and following from existing approaches (as in Joanis et al. (2008)), we use syntactic “slot” features to encode basic argument information about a verb usage. Table 4.1 presents the 14 features used in our representation. The first 12 (up through “PP”) are binary features denoting the presence or absence of the stated syntactic slot, such as an object (OBJ) or a prepositional phrase (PP); the slots are indicated by labels used by the CHILDES dependency parser (Sagae et al., 2007). As in Chapter 3, when a PP is present, the nominal feature PREP denotes the preposition used, and NSLOTS denotes the number of slots present. This representation differs slightly from that of Chapter 3. Firstly, we consider only the slots internal to the verb phrase, ignoring syntactic subjects. Secondly, we introduce a PP slot feature to denote whether any prepositional phrase is present. We observed that there is often little practical distinction among prepositional phrases when marked as PRED, LOC, JCT, or IOBJ (as in Chapter 3), so now when any of those constituents are headed by a preposition, we re-tag them as PP (i.e., they are no longer marked as PRED, LOC, JCT, or IOBJ). This also explains the absence of the IOBJ feature in the current feature set—all IOBJ instances are prepositional phrases. We make the assumption that children at this developmental stage can distinguish various syntactic arguments in the input, but may not yet recognize recurring patterns such as transitive and double-object constructions. The following examples show this representation used with a double-object dative and a prepositional dative, respectively:

(4.3) I sent my mother a letter.

\[ (\text{OBJ, OBJ2, PREP = null, NSLOTS = 2} ) \]

(4.4) I sent a letter to my mother.

\[ (\text{OBJ, PP, PREP = to, NSLOTS = 2} ) \]
### Table 4.1: Slot features.

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJ, OBJ2</td>
<td>Objects</td>
</tr>
<tr>
<td>COMP, XCOMP</td>
<td>Clausal complements</td>
</tr>
<tr>
<td>PRED, CPRED, XPRED</td>
<td>Predicate complements</td>
</tr>
<tr>
<td>LOC</td>
<td>Locatives</td>
</tr>
<tr>
<td>JCT, CJCT, XJCT</td>
<td>Adjuncts</td>
</tr>
<tr>
<td>PP</td>
<td>Prepositional phrases</td>
</tr>
<tr>
<td>PREP</td>
<td>Preposition (nominal value)</td>
</tr>
<tr>
<td>NSLOTS</td>
<td>Number of slots used</td>
</tr>
</tbody>
</table>

#### 4.2.2 Model 1: Argument structure constructions

Like other topic models, the HDP (Teh et al., 2006) is essentially a model of category learning: the model clusters similar items in the input to discover structure. Adopting a usage-based approach to language (e.g., Goldberg, 2006), we view the acquisition of verb argument structure as a category-learning problem. In this view, structured verb knowledge translates well to the hierarchical nature of the model.

Model 1 is a straightforward adaptation of the HDP to verb argument structure, which we will use as a point of comparison for an extended model. Figure 4.1(a) provides an intuitive description of the hierarchical levels of inference in Model 1. At level 1, the lowest level of abstraction, individual verb usages $y_i$ are represented by sets of features as described above.

At level 2, the model clusters similar usages together to form argument structure constructions, where a construction is represented by a set of multinomial distributions, one for each feature. Since the clustering mechanism is nonparametric, we need not specify the total number of constructions to learn. Each of these constructions, denoted...
Figure 4.1: (a) Model 1, a Hierarchical Dirichlet Process applied to learning verb argument structure constructions. (b) Model 2, an extension of Model 1 to learn verb alternation classes. Trans: transitive. DO: double-object construction. PD: prepositional dative.

by its multinomial parameters \( \theta \), probabilistically represents a pattern such as a simple transitive or a prepositional dative. While a construction here encodes only syntactic information, with no semantic elements, the model can be generalized to a combined
syntactic/semantic input representation, as we will see in Chapter 5.

At level 3, a multinomial distribution for each verb ($\pi$) represents the range of constructions that tend to occur with the verb. For example, in Figure 4.1(a), *give* ($\pi_2$) would have a high probability for the double-object dative and prepositional dative constructions ($\theta_2$ and $\theta_3$, respectively), but a low probability for the transitive construction, $\theta_1$. Let $y_{ij}$ denote feature $j$ of usage $i$. Levels 1 through 3 are given by the following:

\[
\begin{align*}
\pi_v & \sim \text{Dirichlet}(\alpha \cdot \beta) \\
\pi & \sim \text{Multinomial}(\pi_v) \\
\theta_{jz_i} & \sim \text{Dirichlet}(1) \\
y_{ij} & \sim \text{Multinomial}(\theta_{jz_i})
\end{align*}
\]

The indicator variable $z_i$ selects a cluster (i.e., a construction, one of the $\theta$s) for usage $i$. Given a verb $v$, this is drawn from a multinomial distribution which includes a small probability of creating a new construction.

The verb-specific distributions $\pi_v$ depend on hyperparameters which encode expectations about constructions in general, across all verbs. They represent acquired knowledge about the likely total number of constructions, which constructions are more likely to occur overall, and so on:

\[
\begin{align*}
\gamma & \sim \text{Exponential}(1) \\
\alpha & \sim \text{Exponential}(1) \\
\beta & \sim \text{Stick}(\gamma)
\end{align*}
\]

As with lower-level parameters, these are influenced by observed structure in the input. $\beta$, drawn from a stick-breaking process (Stick), encodes how many constructions will be used and which constructions are more likely overall. $\alpha$ affects the variability of $\pi_v$.

---

1For binary-valued features, the feature $y_{ij}$ is drawn from a binomial distribution with a beta prior, rather than the multinomial with Dirichlet prior specified here. We continue to discuss the model in terms of the multinomial and Dirichlet as this represents the more general case.
Large values of $\alpha$ push $\pi_v$ closer to $\beta$, the global distribution over constructions, while smaller values encourage more variation among verbs. $\gamma$ affects the total number of constructions; small values of $\gamma$ correspond to fewer constructions. By drawing $\alpha$ and $\gamma$ from an exponential distribution, we give a weak preference for verb-specific behaviour and for solutions with fewer constructions. These preferences are effectively designed into the model; they may be informed by general human category-learning behaviour. For further details of this model, see Teh et al. (2006) or Appendix B.

### 4.2.3 Model 2: Alternation classes

Model 1 acquires argument structure constructions from individual verb usages, and learns how those constructions are used by individual verbs, but it is unable to recognize that certain *kinds* of verbs behave differently than others. Competent language speakers regularly use this kind of information. For example, if a verb occurs in a double-object dative construction, then we should infer that it is also likely to occur in a prepositional dative. We develop a novel extension of the above model to capture this phenomenon by learning clusters of similar verbs.

Recall that we represent a verb by a probability distribution over the constructions in which it may occur. In the example shown in Figure 4.1(a), *give* and *show* both tend to occur with a double-object dative and a prepositional dative, but are less likely to occur as simple transitives. By recognizing the similarity of $\pi_2$ and $\pi_3$, we can create a cluster containing *give*, *show*, and other similar verbs. Figure 4.1(b) presents this intuition in Model 2. We extend Model 1 by introducing a fourth level of abstraction, where we represent clusters of similar verbs. For each verb cluster $c$, we use $\phi_c$ to represent the range of constructions that tend to occur with any of the verbs in that cluster. By serving as a prior on the verb-level parameters $\pi_v$, $\phi_c$ directly influences each verb in the cluster.

The lower levels of this model are the same as those of Model 1. In addition, the verb
representations, \( \pi_v \), depend on the alternation classes in level 4:

\[
\begin{align*}
\phi_{c_v} &\sim \text{Dirichlet}(\alpha_0 \cdot \beta_0) \\
\pi_v &\sim \text{Dirichlet}(\alpha_1 \cdot \phi_{c_v}) \\
z_i &\sim \text{Multinomial}(\pi_v) \\
\theta_{jz_i} &\sim \text{Dirichlet}(1) \\
y_{ij} &\sim \text{Multinomial}(\theta_{jz_i})
\end{align*}
\]

Each verb \( v \) belongs to a cluster of verbs, denoted \( c_v \). Now, \( \pi_v \) depends on \( \phi_{c_v} \), which gives a distribution over constructions for all the verbs in the same cluster.

As before, these parameters themselves depend on top-level hyperparameters:

\[
\begin{align*}
\gamma_0 &\sim \text{Exponential}(1) \\
\alpha_{0,1} &\sim \text{Exponential}(1) \\
\beta_0 &\sim \text{Stick}(\gamma_0)
\end{align*}
\]

These hyperparameters serve similar roles to those in Model 1. \( \beta_0 \) gives a global distribution over all the constructions in use. \( \gamma_0 \) affects the total number of constructions overall. \( \alpha_1 \) affects the variability of a verb compared with its class, and \( \alpha_0 \) affects the variability of verb classes.

To group verbs into alternation classes, we use a mechanism similar to the way we group individual verb usages into constructions. Recall that \( c_v \) acts as an indicator variable, selecting a class for verb \( v \) from the available classes in level 4. This is drawn from a multinomial distribution \( \sigma \) which includes a small probability of creating a new verb class:

\[
\begin{align*}
\gamma_1 &\sim \text{Exponential}(1) \\
\sigma &\sim \text{Stick}(\gamma_1) \\
c_v &\sim \text{Multinomial}(\sigma)
\end{align*}
\]
As with earlier uses of the stick-breaking construction, $\gamma_1$ affects the expected total number of verb classes. This method of clustering verb types is similar to that of Wallach (2008), as I will discuss in Section 4.5. Further details of the stick-breaking construction and our inference method can be found in Appendix B.

### 4.2.4 Parameter estimation

Models 1 and 2, as written, each specify a prior distribution over the complete set of possible parameters to the models (i.e., all possible values for $\theta$, $z$, $\phi$, and so on). We update these distributions using the observed verb usage data, thus obtaining posterior distributions over parameters.

Similarly to our experiments in Chapter 3, we estimate the posterior distributions using Gibbs sampling, a Markov Chain Monte Carlo (MCMC) method (Teh et al., 2006). Model parameters are initially set randomly, then iteratively adjusted according to the observed data. We randomly set each $z_i$ to one of 10 initial constructions, and each $c_v$ to one of 10 verb classes (if applicable). We set the remaining parameters to random values drawn from the distributions specified in the model descriptions. We then iteratively update each model parameter individually by drawing it from a posterior distribution conditioned on the data and all the other parameters in the model. As we iterate through the parameters many times, we collect samples of their values. Over time, this set of samples converges on the posterior distribution—i.e., the model parameters given the observed data. In the experiments, we average over this set of samples to estimate what each model has learned about the input. Please refer to the appendix for details of the MCMC inference.
4.3 Experimental set-up

We use child-directed speech from the Manchester corpus (Theakston et al., 2001), part of the CHILDES database (MacWhinney, 2000). The corpus covers 12 British English-speaking children between the ages of approximately 2 and 3 years. Using CLAN, a software package for analysis of CHILDES corpora, we extract all child-directed utterances containing at least one verb. We parse the utterances with the MEGRASP dependency parser (Sagae et al., 2007), then reserve every second usage for an evaluation dataset, using the remainder for development. As described above, we extract 14 slot features for each verb usage. The datasets corresponding to each child contain between 4,400 and 10,700 usages and between 239 and 479 verb types. All reported results are obtained using the evaluation data.

Due to flaws in the automatic part-of-speech tagging and parsing, the data contains many errors, particularly in ditransitive constructions. We manually correct the portion of the input related to the dative alternation. For each verb in the development set that occurs with at least one prepositional or double-object dative (as given by the automatic parsing), we draw a sample of up to 50 usages. We repair any cases of incorrectly parsed dative constructions, then duplicate the corrected samples as necessary. Since manual annotation is so labour-intensive, we use this same sample to correct the data for corresponding verbs in the evaluation set. We assume that the proportions of various usages are identical for these verbs across the development and evaluation sets.

For each of the 12 children in the input, we run 10 randomly initialized simulations. The parameters appear to converge within 3,000 iterations, so we run each simulation for 5,800 iterations, discarding the first 3,300 as burn-in. We record a sample of the model parameters on every 25th iteration after the burn-in, giving 100 samples per simulation, 1,000 per child. By averaging over these samples, we can examine the models’ behaviour.
4.4 Experiments and analysis of results

We compare the ability of our two models to acquire knowledge about the usage patterns of verbs in the input and generalize that knowledge to new verbs. Firstly, we examine construction preferences in two related classes of verbs. Secondly, we test whether the models use an abstract representation of the dative alternation to help learn new verbs.

4.4.1 Experiment 1. Verb argument preferences

We examine how our models acquire the usage patterns of verbs in the input by looking at verbs that participate in two different alternation patterns. Earlier, we demonstrated the dative alternation in examples (3) and (4). The benefactive alternation is a related pattern, in which verbs alternate between a double-object form and a prepositional benefactive form, as in the following examples:

\[(4.5) \text{John made his friend a sandwich.} \\]
\[
\langle \text{OBJ, OBJ2, PREP = null, NSLOTS = 2} \rangle
\]
\[(4.6) \text{John made a sandwich for his friend.} \\]
\[
\langle \text{OBJ, PP, PREP = for, NSLOTS = 2} \rangle
\]

We consider all verbs involved in the dative and benefactive alternations, as listed by Levin (1993, Sections 2.1 and 2.2). We test three constructions: the prepositional dative (PD); the double-object construction (DO), whether dative or benefactive; and the prepositional benefactive (PB). Using the samples of the model parameters, we estimate the posterior predictive likelihood of each of these frames for each of the verbs in the given classes. For a given test frame $y_0$, using verb $v$, and the observed data $Y$,

\[
P(y_0|Y) = \sum_k P(y_o|k, Y)P(k|v, Y) = \sum_k \prod_j P(y_{0j}|\theta_{jk})P(k|\pi_v)
\]  \hspace{1cm} (4.7)
Figure 4.2: Argument preferences for known dative and benefactive verbs in Models 1 and 2. Shorter bars indicate higher likelihood. The two models show similar behaviour. Dative: verbs listed by Levin (1993) as dative but not benefactive. Benefactive: verbs listed as benefactive but not dative. Both: verbs listed in both classes. PD: prepositional dative. DO: double-object construction. PB: prepositional benefactive.

This likelihood is averaged over all 1,000 samples per child. Since the domain for this probability distribution includes all possible combinations of the features in the input, the computed likelihoods for a given frame can be quite small. Consequently, we report the natural logarithm of these likelihood values and consider the relative differences among frames.

Figure 4.2 shows the behaviour of both models. We average the likelihoods over all 12 children, and over all verbs in the following cases: (a) verbs listed as dative but not benefactive, (b) verbs listed as benefactive but not dative, and (c) verbs in both classes.

---

2This gives an estimate of the relative preferences for a verb’s usage and is a direct measure of the acquired lexicon. Translating this estimate to production would require a model of how discourse and other factors influence dative and benefactive production (e.g., de Marneffe et al., 2011). This is beyond the scope of this thesis.
In both models, both dative and benefactive verbs show a high likelihood for the DO frame, and a somewhat higher likelihood for the appropriate prepositional frame (PD and PB, respectively) than for the inappropriate one (PB and PD, respectively). Verbs that occur in both classes show closer likelihoods for all three frames. The likelihoods of the three prepositional forms are generally similar for two reasons. Firstly, since the syntactic frames differ by only one or two features, a prepositional dative may (for example) be recognized as a partial match to a construction representing a prepositional benefactive. Secondly, the observed usage patterns for these verbs do not necessarily match the classification provided by Levin (1993). For example, a verb listed as strictly dative-alternating may actually be observed in a prepositional benefactive use.

These results suggest that both models can acquire the argument structure preferences of verbs in the input. In this case, the ability of Model 2 to acquire verb alternation classes is not necessary. Both models are able to cluster verb usages into a range of constructions and acquire appropriate usage patterns over a range of verbs. Both models acquire approximately 20 different constructions. Model 2 acquires 35-40 verb classes, depending on the child.

### 4.4.2 Experiment 2. Novel verb generalization

Children as young as three years of age have been shown to use abstract representations of the dative alternation (Conwell & Demuth, 2007). When young children hear a sentence like *I gorped Charlie the duck*, they appear to know that the same meaning can be expressed by saying *I gorped the duck to Charlie*. We test this generalization in our models by presenting a novel verb in one form of the dative and measuring the likelihood of the alternating form.

We test each model by independently presenting it with a novel verb in three different situations: (a) two instances of the prepositional dative, (b) two instances of the double-object dative, or (c) one instance of each. Only in case (c) is the verb explicitly seen
to be alternating. We test the ability to generalize alternation behaviour by comparing the likelihood of the unseen alternating form with an unseen form unrelated to the alternation. The non-alternating frame is the sentential complement (SC) frame, which occurs in 1-1.5% of the input, approximately the same overall frequency as either of the two dative frames. For example, if we train the novel verb using only the PD, yet the DO frame shows a higher likelihood than the unrelated SC frame, then we can say that the model has generalized the dative alternation.

Since the novel verbs are not in the observed data, we must further iterate the Gibbs sampler, using the new data, to obtain the appropriate samples of the verb-level distribution $\pi_v$. For each of the 1,000 parameter samples per child we obtained from the original simulations, we re-initialize the model with the parameters from the sample, add in the novel data for case (a), (b), or (c), then do a further 350 iterations, recording 10 new samples of the model parameters. This gives 10,000 new samples per test case, per child. Using equation (4.7) and the new samples, we estimate the posterior predictive likelihood of each of the three constructions.

Figure 4.3 shows how the ability to acquire verb classes aids generalization. In Model 1, without verb classes, only the frames already seen with the novel verb are highly likely. This means that Model 1 is unable to generalize to a different construction not seen in the observed data. In contrast, Model 2 shows better generalization for the dative alternation. When the novel verb is trained with the prepositional dative, the double-object dative shows a much higher likelihood than the unrelated SC frame. A similar effect occurs with DO-only training: the PD frame is now more likely than the SC frame, although only slightly. Compared with Model 1, both dative frames obtain a higher likelihood across all three training cases, while the SC likelihood remains low. The ability to acquire alternation classes improves the ability to learn both alternating constructions.

---

3For example, *I want you to make me a sandwich*, ⟨OBJ, XCOMP, PREP = null, NSLOTS = 2 ⟩.
Figure 4.3: Generalization of novel dative verbs in Models 1 and 2, under various training conditions. Shorter bars indicate higher likelihood. Novel verbs were trained in three conditions: PD only, two instances of the prepositional dative; DO only, two instances of the double-object construction; Alternating, one instance of each. PD: prepositional dative. DO: double-object construction. SC: sentential complement (unrelated to the dative alternation).

One aspect of our results differs from the behaviour observed in children. Our verb-clustering model is more likely to generalize to the double-object form when trained only on a prepositional form, than the other way around (i.e., generalizing from a DO to a PD). However, three-year-old children seem to be biased to the prepositional form, the opposite effect (Conwell & Demuth, 2007). We suggest that this is a result of our small corpora. High-frequency dative verbs tend to be biased toward the double-object form (Campbell & Tomasello, 2001). However, Gries and Stefanowitsch (2004) show that out of 40 alternating verbs in the larger ICE-GB corpus, 19 are prepositional-biased. This strongly suggests that more low-frequency verbs are prepositional-biased than double-
4.5 Discussion of the model

As described in Section 4.2, the joint verb class model is based on a large body of work in nonparametric topic modelling. In this section, I discuss how our work relates to existing topic models of language, some implications of the underlying assumptions in the model, and the relationship between this model and the work of Alishahi and Stevenson (2008), another recent Bayesian model of verb learning.

4.5.1 Relation to existing topic models

Our model is built on mechanisms common to many probabilistic topic models of language and is similar in structure to various models based on latent Dirichlet allocation (LDA; Blei et al., 2003) and the hierarchical Dirichlet process (HDP; Teh et al., 2006). In particular, the model is most directly similar to the so-called cluster-based topic model of Wallach (2008). As with many topic models, this model is intended to capture the topical structure of documents—that a document is made up of a mixture of topics, and each topic defines a distribution over words in the vocabulary. In addition, the model captures the hypothesis that there are natural clusters of documents with similar distributions of topics. In other words, groups of documents will tend to convey similar topics. Wallach’s model jointly infers both a range of topics and a range of latent clusters of documents. As in LDA, each individual word in a document is assumed to be generated by a topic, and each topic provides a multinomial distribution over the words in the vocabulary. For each document in the corpus, the document’s distribution of topics now depends on a cluster-specific prior—that is, the document-level multinomial over topics is generated.
by a Dirichlet distribution, and there is one of these for each document cluster. Since the number of document clusters is unknown, these clusters are handled by a nonparametric Dirichlet process.

The hierarchical levels of inference in Wallach’s model are analogous to the levels in our verb class model, with certain key differences. At the lowest level in the hierarchy, word tokens in the topic model correspond with verb usages in our model. The difference here is that while a word token is univariate, being generated by a multinomial distribution over all words in the vocabulary, a verb usage is multivariate—we assume that a usage is generated by a product of binomial and multinomial distributions, one for each of the syntactic features listed in Table 4.1.

At the next level, topics in the above model correspond with verb argument structure constructions in our model. While a topic is represented by a distribution over the words in which that topic might be realized in text, an argument structure construction is represented by distributions over the features in which that construction might appear. The main difference here between our model and that of Wallach is that we use a nonparametric mechanism to represent constructions—that is, we need not specify the required number of constructions (or topics) as is necessary in Wallach’s model.

At the next level in the topic model, each document is represented by the distribution over topics for that document. Analogously, each verb is represented by its distribution over constructions. The underlying mechanisms in the two models are fairly similar from here on up. While a document’s topic distribution depends on a document-class prior, in our model the verb’s construction distribution depends on a verb-class prior. In both cases, document classes and verb classes are found by a comparable nonparametric clustering process.

One important distinction between our model and the above topic model lies with the treatment of observations, specifically, where in the hierarchy we place these observations. In the above topic model, observations (individual word tokens) lie exclusively
at the bottom level of the hierarchy. This is in contrast to some other models, such as those based on the nested Chinese restaurant process (Blei et al., 2004), which disperse individual observations through multiple levels of the hierarchy. In our case, all the observed argument slot features lie at the lowest level of the hierarchy, directly comparable to the above model. However, the verb type may be considered to be an aspect of the observation used differently from the set of argument features. Here, it is used to identify a specific distribution over constructions. By comparing verbs with documents, we make an assumption that a verb type is observed in the same way as one “observes” the identity of a document in a typical topic model—that is, for a given word token, we assume we know which document contains that token. Likewise, for a given verb usage, we assume that the child can accurately identify the verb to which this usage applies.

We recognize that this assumption does not necessarily hold in general, since there is a degree of uncertainty regarding the verb itself. Since the scope of the model here is to estimate the overall behaviour of a given, known verb, we leave this as an extension for future work.

### 4.5.2 Sequential data and exchangeability

As with many other topic models of language, our model employs a simplifying assumption that individual observations are exchangeable, that is, that the model’s representation of verb knowledge is insensitive to the order of verb usages in the input. This assumption arises in two key ways in our model: firstly, in the probabilistic representation of verbs and their argument structures; secondly, in the method of inference used to estimate the model’s posterior probability distributions given the input.

With respect to the first effect of exchangeability, let us consider how this arises in typical applications of such topic models. In many models, exchangeability arises as the so-called bag-of-words assumption, whereby the order of words within a document is ignored. Since information about word order is ignored, much of the associated syntactic
information (in order-dependent languages such as English) is also ignored.

Our model is not subject to the same issues. Recall that in a typical topic modelling application (e.g., the analogy described in the previous subsection), a single observation is a word token; the exchangeability assumption effectively mixes up the order of words within a document. In our case, a single observation is a complete verb usage—a multivariate structure that incorporates syntactic information about verb arguments. Exchangeability now applies to the order of verb usages. We assume that the likelihood of a particular construction for a verb is independent of the sequential order of other usages of the verb. This assumption simplifies the representation and inference in the model, while allowing us to capture a large amount of relevant verb learning behaviour.

While we are able to retain much syntactic information with our representation, the assumption of exchangeability still results in a limitation of the model. This assumption does not hold in general—verb usages are not completely independent of their order. For example, adults and children are significantly more likely to produce a particular form of the dative alternation (either the double-object or the prepositional form) if they have recently heard or produced that same form (Thothathiri & Snedeker, 2008; Bresnan et al., 2007; de Marneffe et al., 2011). This form of priming in an alternation even happens across distinct verb types. In other words, hearing someone say *Give Mary some ice cream* would make me more likely to say *Show John the bill* (versus *Show the bill to John*) than if the person had instead said *Give some ice cream to Mary*.

One possible way to lift this restriction, and to capture the dependency above, would be to introduce a Markov chain in the sequence of latent constructions $\theta$ generating verb usages. That is, the likelihood that a given verb usage $y_i$ is generated by construction $k$ (i.e., that $z_i = k$) could also depend on the constructions used for a few preceding usages: $z_{i-1}, z_{i-2}$, and so on. This is a similar approach that has been taken to alleviate the bag-of-words assumption in other topic models (e.g., Griffiths et al., 2005).

The second way exchangeability arises in our model is in the inference method used
to estimate model parameters given the observed data. As detailed in Appendix B, our Gibbs sampler makes use of the exchangeability assumption in the way it samples construction membership—sampling the most likely constructions to have generated each of the verb usages in the input. This method requires that we estimate model parameters given the input in a batch process: we collect all relevant training data, then estimate model parameters all at once.

Children, however, do not have the luxury of waiting to hear all relevant input before building a lexicon. Rather, they must acquire knowledge incrementally, having access to their in-progress abstract representations at every step of the way. The order of children’s acquisition of various concepts and phenomena is a significant part of developmental research, and ultimately computational models should provide ways to investigate such effects.

The batch-processing inference mechanisms we use in this thesis can provide coarse-grained insight into order-of-acquisition effects in language development. To compare the model’s knowledge state with that of children at specific points in development, we can apply the model to different-sized sections of age-appropriate input—effectively, by breaking up one large batch of input into a series of smaller batches. This is similar to the approach taken in Chapter 3 to investigate the order of acquisition of the senses of get. Finer-grained observations of developmental trajectories would most likely require an inference algorithm that steps through the input incrementally, such as the category-learning model proposed by Anderson (1991), used by Alishahi and Stevenson (2008) to acquire verb argument structure, and also used by Parisien et al. (2008) to acquire syntactic categories. Another possible MCMC approach would be to develop a particle filtering algorithm for this model (akin to that of Sanborn et al., 2006). The development of such an incremental approach is beyond the scope of this thesis and is thus proposed for future research.
4.5.3 Noise

Since we apply our model to naturalistic child-directed speech, using automatic methods to extract input features, the input to the model is subject to a significant amount of noise. As mentioned in Section 3.2.1, we argue that this noise is reasonable, as it captures the complexity of the input children receive. There are two main sources of noise in the input. The first source of noise is infelicities in the child-directed speech in the original corpora—as is typical in natural speech, parents frequently speak in incomplete or ungrammatical sentences. The second source of noise is a result of the automatic tools for extracting features of the input. In particular, the MEGRASP dependency parser (Sagae et al., 2007) used to extract syntactic argument slots is subject to a significant amount of error. We argue that this is reasonable given the difficulty that children experience in distinguishing such arguments. We note that this source of noise is not uniform—personal experience suggests that given a particular verb argument structure (e.g., the double-object construction), the usages of one verb (e.g., give) may be subject to a higher error rate than those of another verb (e.g., show).

For the purpose of aiding generalization, noisy input may actually be beneficial. By introducing observations of low-likelihood events, noise increases the variability of the observations, flattening out some of the relevant probability distributions in the model. Since generalization depends on predictions of unseen events, noise can thus increase the likelihood that a model will generalize beyond its input.

4.5.4 Relation to Alishahi and Stevenson (2008)

As discussed in Section 4.1, the verb class model is closely related to the argument structure acquisition model of Alishahi and Stevenson (2008). Indeed, their model forms a theoretical and practical foundation for the work in this thesis. I do not attempt to refute the results of these authors; rather, I aim to build on their work. Here, I clarify
some of the connections between this thesis and the work of Alishahi and Stevenson (2008).

There are two key aspects of the model of Alishahi and Stevenson that are central to this thesis. Firstly, the authors proposed that verb argument structure constructions could be represented as probabilistic associations between form and meaning. Moreover, these probabilistic constructions could be acquired from child-directed speech by clustering individual verb usages on the basis of their similarity. This has provided a valuable connection between abstract linguistic knowledge and usage-level input. In this chapter, we use the same notion of clustering verb usages to acquire probabilistic representations of the syntactic aspects of argument structure, extending these abstractions to capture high-level knowledge about verb alternations. In Chapter 5, we augment the representation to capture associations between form and meaning, showing how this knowledge impacts higher-level generalization in verb learning. Secondly, the probabilistic representation and inference in the model of Alishahi and Stevenson is based on the category-learning algorithm proposed by Anderson (1991). Given a particular parameter setting, Anderson’s model is directly equivalent to a DPMM (Neal, 2000), the model used in Chapter 3. Moreover, this nonparametric representation and inference forms the basis of the models in this chapter and in Chapter 5. Due to the similarity of the underlying representational framework to that in our own work, we consider the observations of this thesis to be complementary to those of Alishahi and Stevenson.

4.6 Summary of contributions

In this chapter, we showed how verb alternation classes contribute to generalization in verb learning. We developed a hierarchical Bayesian model, Model 2, that is capable of acquiring knowledge of verb argument structure at multiple levels of inference simultaneously. We demonstrated this using the wide range of verbs and constructions contained
in a corpus of naturalistic child-directed speech.

By clustering individual verb usages, both of our models acquire a variety of argument structure constructions and learn their patterns of use over hundreds of verbs. Furthermore, Model 2 learns groups of verbs that occur with similar usage patterns. Using the dative alternation as a key example, we demonstrated how this knowledge of alternation classes can be generalized to novel verbs, as observed in the behaviour of children and adults. This verb class model can acquire and apply this knowledge without any prior expectation of which constructions and alternations may or may not be relevant. To our knowledge, this is the first computational model that jointly acquires both verb argument structure and verb classes from a naturalistic corpus.

In contrast to previous analyses of the dative alternation (Perfors et al., 2010; de Marneffe et al., 2011), we demonstrated its acquisition in the context of many other constructions, verbs, and alternations. Despite the low frequency of the participating constructions, our model successfully acquires the dative alternation. This is a strong endorsement of hierarchical Bayesian models of language acquisition.

This approach potentially unifies differing theoretical positions in language acquisition. By simultaneously learning at multiple levels of abstraction, our model combines a usage-based representation of language, as proposed by Tomasello (2003), with weak abstract representations similar to those suggested by Fisher (2002a). Other usage-based Bayesian models, such as that of Alishahi and Stevenson (2008), offer a similar opportunity, although our model develops higher-level abstractions regarding the structured knowledge of verbs.

One of the key features of usage-based constructions is that they couple form to meaning (Goldberg, 2006). Moreover, abstract syntactic representations appear to influence semantics in verb learning, and vice-versa. Two-year-old children have been shown to use the alternation structure of a novel verb to predict aspects of its meaning (Naigles, 1996; Scott & Fisher, 2009), and somewhat older children are able to use the meaning
of a novel verb to predict its range of acceptable syntactic structures (Ambridge et al., 2011; Kline & Demuth, 2010). In the next chapter, we augment our model’s input with semantic properties to investigate this interaction between syntax and semantics in verb alternations. In doing so, we demonstrate that syntactic and semantic associations in argument structure, acquired from the input, can be robust enough to support high-level generalizations between form and meaning in verb learning.
Chapter 5

Generalizing between form and meaning

To communicate effectively, children must acquire the kinds of knowledge that allow them to be productive with language. As children learn how to generalize their linguistic knowledge to new situations, the emergence of these generalizations signals important changes in children’s representations of language. To understand this developmental process, we must investigate how children can find the abstract representations that lie at the core of these generalizations. Moreover, we need to consider what kinds of inferential mechanisms might allow these abstractions to actually guide generalization.

As we discussed in Chapter 2, several lines of research suggest that, in many situations, children’s ability to generalize is governed by strong regularities between form and meaning. Much of these regularities occur within alternations in verb argument structure, in which verbs show different patterns in how they can express their semantic arguments in syntactic forms. A good example of this occurs with the English verb break, which participates in the causative/inchoative alternation:

\[(5.1) \text{John}_{\text{Agent}} \text{ broke the window}_{\text{Patient}}. / \text{The window}_{\text{Patient}} \text{ broke.}\]

In this alternation, the object of the transitive (the Patient, window) becomes the subject
of the intransitive. Contrast this with the verb *laugh*, which also occurs in both transitive and intransitive forms, but with differing semantic roles. In this case, the intransitive is much more frequent than the transitive:

(5.2) Jane\textsubscript{Agent} laughed her glee\textsubscript{Theme}. / Jane\textsubscript{Agent} laughed.

Such patterns are not accidental: the pattern with *break* is common with change-of-state verbs (such as *freeze*, *split*, and *dry*), while that of *laugh* is typical of expression verbs (like *cry*, *snort*, and *giggle*). Alternation patterns thus capture a connection between the semantics of verbs and their syntactic expression. As discussed in previous chapters, this reflects a possible cognitive class structure of verbs (Levin, 1993).

Moreover, such form–meaning regularities have been shown to influence language learning. Two-year-old children can use the alternation structure of a novel verb to predict aspects of its meaning (Naigles, 1996; Scott & Fisher, 2009), and somewhat older children can also use aspects of a verb’s meaning to predict its range of acceptable syntactic structures (Ambridge et al., 2011; Kline & Demuth, 2010). This kind of inference in language acquisition appears to involve the interaction of many complex factors, including frequency, verb meaning, and animacy of the arguments. Human and computational experiments have clearly demonstrated the role of statistical regularities over such factors in guiding generalization behaviour (e.g., Merlo & Stevenson, 2001; Scott & Fisher, 2009; Perfors et al., 2010). The next step is a computational model of child language acquisition that models such inferences over verb alternations in the face of noisy, real-world data such as children receive.

In the previous chapter, we developed a hierarchical Bayesian model that learns abstract knowledge of verb argument structure and verb classes from naturalistic child-directed speech. We initially demonstrated this model by using syntactic features of the input, to show what could be learned about verb alternations by attending only to syntactic aspects of argument structure. However, the model did not directly address semantic generalization. In this chapter, previously published in large part as Parisien and
Stevenson (2011), we extend this model to capture form–meaning regularities relevant to alternation patterns. We show that the complex probabilistic abstractions acquired by the model are robust enough to capture key behaviours of children and adults in generalizing over verb alternation knowledge. Specifically, we simulate key results from the above human experiments. We show that the model can use the alternation structure of a novel verb to predict aspects of its meaning (as in Naigles, 1996; Scott & Fisher, 2009), and that the model can use a verb’s meaning to predict its range of use (Ambridge et al., 2011; Kline & Demuth, 2010). By demonstrating specific interactions among various factors in verb learning, this computational framework permits us to make key predictions regarding developmental behaviour.

We argue that this kind of probabilistic representation is critical for learning about alternations, since it gives an explicit role for input frequency and allows detailed interactions between frequency and the cooccurrence of various form and meaning features. Moreover, by using verb classes to capture general tendencies over alternations in the data, this representation alleviates the effect of noise and uncertainty inherent in real-world usages of verbs, which show individual variation in their adherence to typical alternation patterns. These properties make this a useful framework for investigating the predictions that arise from the many interacting factors in verb learning.

The remainder of this chapter proceeds as follows. Section 5.1 describes the semantic extensions to the model that permit it to capture generalizations between form and meaning in verb knowledge. Sections 5.2 and 5.3 describe two experiments that simulate key behaviours in child development, showing how the model can be used to further investigate acquisition and generalization of argument structure and alternations in children. Lastly, Section 5.4 summarizes the contributions of this chapter.
5.1 Model description

We first present the feature representation of the verb usages that serve as input to the model, and then describe how the hierarchical representations of alternation classes capture syntactic and semantic generalizations over verbs.

5.1.1 Representation of Verb Usages

Our representation of individual verb usages comprises both syntactic and semantic information. For the syntactic side, we use the representation from Chapter 4, which includes 14 features for the number and type of syntactic arguments occurring with a verb. The arguments are recorded individually, under the assumption that children at this developmental stage can identify these various syntactic arguments in the input, without necessarily being able to keep track of full subcategorization frames (a more difficult task).

We extend the representation to add a further 15 binary features which capture general semantic information about a verb usage. The first of these features denotes the animacy of the syntactic subject, a method previously used to help distinguish the Agent from other roles in subject position (e.g., Merlo & Stevenson, 2001; Joanis et al., 2008). The next 14 features denote the presence or absence of various coarse-grained semantic properties concerning the event described by the verb. We use general features (not tied to specific verbs or classes) that capture a wide range of verb semantic characteristics, thereby enabling the model to distinguish important aspects of verb semantics discussed in the acquisition literature. While the behaviour of the model is not dependent on any specific set of features, in this work we adopt the following semantic predicates that have been used in the VerbNet verb classification (Kipper-Schuler, 2005): cause, exist, motion, direction, contact, force, has-possession, perceive, experience, expression, disappear, emit, change-state, and result.
Figure 5.1: Structure of the model. The acquired constructions provide probabilistic associations between syntactic and semantic aspects of argument structure.

The following examples show this representation. (Binary features with a value of 1 are listed, along with the value of non-binary features.)

(5.3) John broke the window.

\[ \{ \text{OBJ, NUMSLOTS = 2, SUBJ = animate,} \]
\[ \text{CAUSE, CONTACT, CHANGE-STATE, RESULT } \} \]

(5.4) The window broke.

\[ \{ \text{NUMSLOTS = 1, SUBJ = inanimate,} \]
\[ \text{CHANGE-STATE, RESULT } \} \]

5.1.2 The hierarchical model of verb knowledge

In the previous chapter, we introduced the mathematical structure of the model and explained how the hierarchical levels of representation in the model correspond to distinct levels of abstraction in verb knowledge. Here, we offer a high-level review of this representation as it relates to the interaction between verb alternation classes and verb semantics. Figure 5.1 provides an intuitive description of the distinct levels of inference in the model. This is the same model as shown in Figure 4.1b, using examples corresponding to the experiments in this chapter to show the relationship between syntactic
and semantic information in the model.

At level 1, the lowest level of abstraction in the hierarchy, individual verb usages are represented by sets of syntactic and semantic features as described above. At level 2, the model probabilistically groups similar verb usages into clusters. This set of clusters captures a range of argument structure constructions, where each of these constructions is represented by a set of probability distributions over the syntactic and semantic features in the input. In this way, the model acquires probabilistic associations between form and meaning, a central notion in construction grammar and usage-based language acquisition (Langacker, 2000; Goldberg, 2006). As in all our models, we need not specify the total number of constructions to learn; the model itself selects an appropriate set of constructions to represent the input.

In level 3, for each verb in the input, we estimate a distribution over the range of possible argument structure constructions. This gives a general pattern of usage for each verb in the lexicon. For example, in Figure 5.1, break would have a high probability for at least two constructions: the transitive change-of-state construction (*John broke the window*) and an intransitive form (*The window broke*). A key benefit of this kind of representation is that it can distinguish alternative constructions by their degree of entrenchment. While it is possible to use a verb like laugh transitively (*Jane laughed her glee*), it is far more likely to be used as an intransitive. The intransitive form of laugh should be more entrenched in the lexicon, and should have a greater effect on generalization patterns for laugh and other verbs of expression.

Level 4 of the hierarchy allows the model to acquire classes of syntactically and semantically similar verbs. The model groups together verbs with similar patterns of argument structure use—precisely the probability distributions acquired in level 3. Each one of the verb classes in level 4 is represented by another distribution over argument structure constructions, but this time accounting for the patterns of *all* of the verbs in the class as a group.
These levels in the model—of abstractions over verb usages—are central to its ability to generalize verb knowledge beyond the data explicitly seen in the input. Each level in the hierarchy provides a more general form of knowledge that can be used to make predictions about the level below it, so that all levels play a role in generalization. In this way, we can predict the usage patterns of a relatively infrequent verb like *rend* using knowledge of similar verbs like *break*, *split*, and *crack*. As we discuss below, these generalizations allow the model to predict syntactic and semantic aspects of novel verbs, capturing important aspects of child behaviour.

### 5.2 Experimental set-up

We use the Thomas corpus, a longitudinal study of a British English-speaking boy from 2 to 5 years of age (Lieven et al., 2009), part of the CHILDES database (MacWhinney, 2000). This corpus is larger and more consistent than the Manchester corpus used in Chapter 4. The Thomas corpus was not available when the research of Chapter 4 was conducted. Our input includes all child-directed utterances from this corpus that have at least one verb, using every second usage for development data and the rest for evaluation. The evaluation dataset contains 170,076 verb usages and 1,393 verb types. All reported results are obtained from evaluation data.

The 14 syntactic features for each verb usage are extracted using the parser of Sagae et al. (2007). We manually annotate as animate or inanimate all 4,213 noun phrases that occur as subjects in the input. We estimate the 14 event semantic features for each usage using VerbNet (Kipper-Schuler, 2005), as follows. We look up all the argument frames in VerbNet (over all senses/classes of the verb) that are compatible with the syntactic frame of the current usage, and extract all the semantic features associated with each such frame. We determine compatible VerbNet frames by comparing syntactic slot features: the VerbNet frame must contain as many postverbal NP arguments as there
are objects in the usage and as many S arguments as there are predicate complements in
the usage (i.e., COMP or XCOMP). There must be at least as many PPs in the usage as
there are specified in the VerbNet frame. The extracted set of VerbNet semantic features
combines those found over all senses/classes of the verb. Following this extraction, for
each of the 14 binary semantic features described in Section 5.1.1, the feature for that
usage is marked as True if it is contained in the extracted set. This procedure results in
a very noisy representation of the semantics of a verb usage. In particular, because the
features for a usage are drawn from all possible senses of the verb in that frame (and
not just its intended sense in the usage), the semantics includes features from VerbNet
classes that are irrelevant to that usage. Thus, while we use VerbNet to enable us to
automatically determine a reasonable set of semantic features, this process does not
simply build perfect information about the VerbNet classes of the verb usages into our
input. Moreover, the automatic extraction of both the syntactic and semantic features
yields a noisy representation that is reasonable given the capabilities of young children
in determining such properties.

As described here and in Chapter 4, the model is defined by the parameters of a
set of probability distributions representing each level of abstraction. As before, we em-
ploy a Gibbs sampling method based on the HDP implementation of Teh et al. (2006).
This is an iterative process that results in a large number of samples from the poste-
rrior distribution—i.e., the model parameters given the observed data. On development
data, the parameters always converge within 3,000 iterations. We perform 10 randomly
initialized MCMC simulations on the evaluation data, running each simulation for 5,550
iterations, discarding the first 3,050 as burn-in. We record a sample of the model pa-
rameters on every 25th iteration after the burn-in, giving 100 samples per simulation, for
1,000 in total. In the experiments, we average over this set of samples to estimate what
the model has learned about the input.

In the simulations, the model acquires approximately 100 argument structure con-
instructions and 90–100 verb classes. Particularly in the smaller classes, low frequency verbs tend to be placed in several different classes over different parameter samples, which is a reflection of the uncertainty in classifying infrequent verbs.

5.3 Experiments and analysis of results

Using its abstract knowledge, the model exhibits two important forms of syntactic and semantic generalization. Firstly, we show how the model can use distributional cues in the alternation structure of a novel verb to infer previously unobserved aspects of its meaning. Secondly, we demonstrate that the model uses the semantic class of a novel verb to appropriately constrain its expected alternation behaviour.

5.3.1 Experiment 1. From alternations to verb meaning

Two-year-old children have been shown to use the alternation structure of a novel verb to infer aspects of the verb’s meaning (Naigles, 1996; Scott & Fisher, 2009). For example, in Scott and Fisher (2009), children first heard a dialogue (audio-only) containing a novel verb used with one of two different alternation patterns—i.e., two combinations of transitive and intransitive usages with varying animacy of the arguments. In one pattern, the novel verb was used in a causal alternation (Matt dacked the pillow / It dacked), while in another, the verb was used in an unspecified-object alternation (Matt dacked the pillow / He dacked). The children were then shown two videos with two different events—one video showed a causal event (e.g., a girl pushing on a boy’s shoulders to make him crouch), while the other showed a contact activity (e.g., the girl brushing the boy with a feather duster). Children were asked to choose the event matching the just-heard novel verb. Although the children were not shown a depiction of the novel verb when they heard the dialogue, they were able to map the verb to the semantically-appropriate visual scene based solely on its alternation pattern. If they heard a causal alternation, they tended
to pick the causal event. If they heard the unspecified-object alternation, they tended to pick the contact activity.

**Experimental design**

We test our model’s ability to generalize in this way from alternation patterns to verb semantics, as follows. We present a novel verb to the model in a particular alternation pattern (*i.e.*, two different syntactic usages that correspond to an alternation), but without any event semantics (*i.e.*, the 14 semantic features corresponding to general verb semantics are left blank). This corresponds to a child hearing these usages of the novel verb without seeing a corresponding scene. We include subject animacy in the presented usages, because although it is a semantic feature, it can be reflected in pronoun use in a dialog, as in Scott and Fisher (2009), even if no scene is observed. We then compare the likelihood of two possible events paired with the verb, one much more compatible with a verb class displaying that alternation, and one much less. The event that is deemed more likely by the model should be the one with the semantic features that match those expected for a verb with the given alternation behaviour. In other words, the model should use the alternation pattern of a novel verb to choose a scene with appropriately matching event semantics.

We use novel verbs comparable to two English verb classes that differ in overall alternation patterns: change of state (*e.g.*, break, freeze, dry) and nonverbal expression (*e.g.*, laugh, giggle, cry). Both types of verbs occur in both transitive and intransitive usages (see Examples 1 and 2), but with differences in two important aspects. First, they differ in the relative frequency of occurrences in these frames. The change-of-state verbs overall occur roughly equally in each frame, while the expression verbs occur predominantly in the intransitive. Second, because of the differing roles taken by their subjects, they have different patterns of subject animacy (since Agents tend to be animate more than Patients). Change-of-state verbs have animate subjects about 70% of the time.
Table 5.1: Training conditions for the novel verb in Experiment 1, each with 24 total usages. The four conditions explore a possible interaction between syntactic frame (Alt or Intrans) and subject animacy (AnimCS or AnimEX), two possible cues to verb meaning. Alt-AnimCS corresponds to a canonical change-of-state verb in terms of syntactic frame and subject animacy. Intrans-AnimEX corresponds to a canonical expression verb. Alt-AnimEX and Intrans-AnimCS are crossed conditions, drawing one cue from each of the two semantic verb classes.

<table>
<thead>
<tr>
<th>Experimental condition</th>
<th>Transitive</th>
<th>Intransitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alt-AnimCS</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Alt-AnimEX</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Intrans-AnimCS</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Intrans-AnimEX</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

in the transitive and 50% in the intransitive, while expression verbs have animate subjects about 80% of the time in both frames.

We present the model with usages of a novel verb in four different conditions (independently), each having 24 usages in one of four alternation patterns, shown in Table 5.1. Recall that there are two possible cues in the dialogue that could be used to distinguish a change-of-state verb from an expression verb: the syntactic frames (either alternating or predominantly intransitive) and the animacy of the intransitive subject (either half inanimate or predominantly animate). We use training conditions such that the novel verbs are shown in alternation patterns with a varying proportion of transitive/intransitive usage, and a varying proportion of animate subjects, with each variation reflecting idealized usages of the two types of verbs. This allows us to examine a possible interaction between the syntactic frame patterns and subject animacy patterns. The two-part nam-
ing of the conditions in Table 5.1 corresponds to each of these two cues. The first term indicates that the syntactic frame pattern either alternates freely between a transitive and intransitive usage (Alt) or is predominantly intransitive (Intrans). The second term indicates the animacy of subjects in intransitive frames, reflecting idealized proportions of a change-of-state verb (AnimCS) or an expression verb (AnimEX). The first condition, Alt-AnimCS, corresponds to what we see for a canonical change-of-state verb. The last condition, Intrans-AnimEX, corresponds to a canonical expression verb. The other two are crossed conditions, drawing one cue from each of the two semantic verb classes. In all conditions, all of the event semantic features of the presented usages are left unspecified, corresponding to the child hearing a dialogue with a particular alternation pattern and with no accompanying depiction of the verb.

For each of these four training conditions, we then present the model with two test frames. Both are intransitive with an animate subject (consistent with a novel verb of either semantic class), and one has the semantics of a change-of-state verb, while the other has the semantics of an expression verb, as follows:

\((5.5) \langle \text{SUBJ = animate, CHANGE-STATE, RESULT} \rangle\)

\((5.6) \langle \text{SUBJ = animate, EXPRESSION} \rangle\)

We then compare the preference in the model for each of these two frames, to see whether the model can infer appropriate semantics from an alternation pattern.

**Experimental results**

Given the observed usages of the novel verb, \(Y_{nov}\), in one of the four conditions, Alt-AnimCS, Alt-AnimEX, Intrans-AnimCS, Intrans-AnimEX, we estimate the likelihood of each test frame \(y_{test}\) using the acquired argument structure constructions \(k\) and verb classes \(c\):

\[
P(y_{test}|Y_{nov}) = \sum_k \sum_c P(y_{test}|k)P(k|c)P(c|Y_{nov})
\] (5.7)
Figure 5.2: Using the alternation and animacy patterns of a novel verb to infer meaning. The plot shows the percentage preference for the scene with change-of-state semantics. From left to right, the conditions from Table 5.1 are: Alt-AnimCS, Alt-AnimEX, Intrans-AnimCS, Intrans-AnimEX. There is a clear interaction between the two dialogue cues: both syntactic frame and subject animacy must correspond to an expression verb in order for the model to prefer the expression scene.

This estimate considers how likely the test frame would be if the novel verb happened to be a member of each class $c$, weighted by the probability that $c$ is the true class for the observed pattern $Y_{nov}$. We normalize these likelihoods over the two test frames, and average the preference over all 1,000 samples from the simulation.

Figure 5.2 shows the percentage preference in the model for the test frame with the change-of-state semantics. A high value indicates that the model has a preference for a change-of-state interpretation for the novel verb, while a low value indicates an expression interpretation. When the input alternates between transitive and intransitive frames, there is a strong preference for the change-of-state scene. There is a small effect of animacy here, such that the animacy pattern of an expression verb reduces this preference slightly. When the input consists entirely of intransitive usages, we see preferences in line with the predictions of the animacy feature: there is a preference for the change-of-state
scene in the AnimCS animacy condition, and for the expression scene given AnimEX animacy.

There is thus a strong interaction between the alternation cues and the animacy cues. When animacy reflects a change-of-state verb (the black bars in Figure 5.2), the alternation has no effect on the preference. When animacy reflects an expression verb (the white bars in Figure 5.2), the alternation pattern has a strong effect. Verb usage patterns in the corpus provide a strong bias for a change-of-state interpretation of a novel verb, and the model requires two strong cues (frequent intransitives as well as frequent animate subjects) in order to pull its interpretation in favour of an expression verb. These results show that the model can use the distributional information carried over multiple syntactic frames to help infer the meaning of a novel verb. Moreover, this shows how two distinct features interact to guide generalization behaviour.

Scott and Fisher (2009) discuss possible mechanisms children might use in making this generalization. They consider a category-mediated process, similar in principle to our model, as well as a direct inference process, by which children directly employ distributional cues to interpret the novel verb, without recourse to a previously learned class. The experiment above corresponds to a category-mediated process, since the preference in Equation 5.7 draws on the class knowledge in the model. Using the estimated model parameters, we repeat the above experiment instead using one possible method of direct inference. Rather than measuring the scene preference by comparing the novel verb to each verb class, as in Equation 5.7, we instead compare the novel verb directly against each of the known verbs from the input. By doing so, in all four training conditions, we observe a 96-98% preference for the change of state scene, with no clear effect of syntactic frame or animacy use. This is a result of drawing inferences over 1,393 verbs, where noise in the data is compounded over such a large number of comparisons. By using verb classes to capture general tendencies in the data, a category-mediated model helps to alleviate the effect of noise, providing better inference in generalization.
5.3.2 Experiment 2. From verb meaning to alternations

The previous experiment considered cases where the alternation structure of a novel verb can help determine the verb’s meaning. The reverse can also be true: information about a novel verb’s semantic class constrains adults’ and children’s expectations concerning the syntactic structures that can be used with the verb (Ambridge et al., 2011; Kline & Demuth, 2010). For example, in the Ambridge et al. experiments, subjects were taught a novel verb that was used only in intransitive frames, while also being shown the meaning of the verb. They were then asked to rate the verb in a transitive usage. Subjects were more likely to rate the transitive use of the verb as acceptable if its semantics matched a class of verbs which display a transitive/intransitive alternation, than if the semantics matched a class that is predominantly intransitive. They did so even though both types of novel verbs had been seen in equivalent intransitive-only usages. That is, the semantic class of the novel verb constrains its generalization to a previously unobserved syntactic usage. Here, we show how verb semantics and entrenchment can similarly be used to constrain generalization in our model.

Ambridge et al. also expected to find an effect of input frequency on the generalization, such that the acceptability of a transitive frame would be lower for a high-frequency novel verb (i.e., presented 24 times) than for a low-frequency novel verb (presented only 8 times). That is, as the intransitive use becomes more entrenched, it would more strongly constrain the use of the transitive. The authors did not observe such an effect with novel verbs, although they did find this effect with known verbs (which they also had subjects rate in various usages). The authors do note that the lack of an observed general effect of frequency may have been a result of the limited range of chosen frequencies—the chosen range may have been too narrow to capture the effect.
**Experimental design**

We simulate this experiment by presenting our model with novel verbs comparable in meaning to five different semantic classes, the same classes used by Ambridge et al. Verbs in the first three classes occur freely in the intransitive, but are much less likely to be used in the transitive: disappearance (e.g., disappear, die), directed motion (fall, tumble), and nonverbal expression (laugh, giggle). The other two classes are likely to alternate between transitive and intransitive forms: manner of motion (roll, spin) and change of state (break, split).

We replicate the two frequency conditions used by Ambridge et al. and also introduce a third, lower-frequency, condition. Based on our intuition that the Bayesian model is more sensitive to small amounts of data than a child would be, we use a set of only 2 frames as the lowest-frequency case.\(^1\) Thus, we present the model with 2, 8 or 24 intransitive frames of a novel verb, coupled with semantic values from one of the following five verb class conditions:

<table>
<thead>
<tr>
<th>Non-alternating classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disappearance:</td>
</tr>
<tr>
<td>Directed motion:</td>
</tr>
<tr>
<td>Nonverbal expression:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alternating classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manner of motion:</td>
</tr>
<tr>
<td>Change of state:</td>
</tr>
</tbody>
</table>

As with the previous experiment, we set the proportion of frames with animate versus inanimate subjects in accord with the proportions for these classes in our corpus data.\(^2\)

---

\(^1\)We note that since the model’s sensitivity differs from that of children, the 8- and 24-frame conditions do not necessarily match those studied in the original experiment.

\(^2\)It is unclear what proportions of subject animacy were used by Ambridge et al. (2011).
Figure 5.3: Using the meaning of a novel verb, shown only as an intransitive, to constrain the likelihood of the transitive.

**Experimental results**

Given the training conditions of the novel verb—i.e., a set of intransitive usages with each of 5 possible semantics (as indicated above), at a given frequency level (2, 8, 24)—we use Equation 5.7 to measure the likelihood of an intransitive and a transitive test frame. (The test frames each have an animate subject, since those are more frequent overall.) Since the test frame likelihoods produced by this method cannot be directly compared with acceptability ratings (as in the human experiments), we instead report the likelihood of the transitive frame relative to that of the intransitive. That is, we divide the transitive likelihood by the intransitive likelihood and report this ratio in Figure 5.3.

Firstly, Ambridge et al. (2011) observed that when the meaning of the novel verb matched a class of alternating verbs, participants rated transitive uses as more acceptable than if the meaning matched a non-alternating class. In our results, transitive verb usages are more acceptable in the manner-of-motion and change-of-state conditions than in the other three cases. That is, when the novel verb has a meaning similar to a class of alternating verbs, it is more expected to alternate, despite only ever being seen in the
intransitive form. The model uses information about the semantic class of the novel verb to appropriately constrain generalization patterns.

Secondly, Ambridge et al. expected to find an effect of input frequency on the generalization, but they only observed the effect with known verbs, not with novel verbs. We do see the effect in our results on novel verbs, in some conditions. Specifically, for both of the alternating classes and for one of the non-alternating classes, the likelihood of the transitive decreases as the input frequency increases. The novel verb becomes increasingly entrenched as an intransitive-only verb, even though this may conflict with the semantic cues (i.e., in the case of the novel verbs from the alternating classes). These results show how the model can be used to investigate the interaction of multiple factors in verb learning: semantic cues still have an effect at higher frequencies, but the effect is tempered by the increasing frequency of the observed frame.

5.4 Summary of contributions

In this chapter, we showed how abstract knowledge of verb argument structure and verb alternation classes contributes to syntactic and semantic generalization in verb learning. We extended the hierarchical Bayesian model of verb learning to capture form–meaning regularities in argument structure, and showed that the complex probabilistic abstractions captured by the model are robust enough to drive realistic generalizations of verb alternation knowledge.

Our model is capable of using distributional information carried over multiple syntactic frames to infer aspects of the meaning of a novel verb, a generalization effect observed in children (Naigles, 1996; Scott & Fisher, 2009). We showed that by capturing general tendencies of verb use, probabilistic representations of verb classes help to alleviate the effect of noise characteristic of real-world data.

The model is also capable of using the meaning of a novel verb to constrain alternation
patterns, an effect discussed by Pinker (1989) and demonstrated recently by Ambridge et al. (2011) and Kline and Demuth (2010). Moreover, we showed that as the frequency of the novel verb increases in training, the entrenchment of the observed pattern further constrains generalization, an important factor in usage-based approaches to language acquisition.

To our knowledge, this is the first computational model of verb learning from real-world data that demonstrates the use of acquired class-level knowledge to show both syntactic and semantic generalization effects. The probabilistic nature of the representation is robust to the noise and uncertainty inherent in child-directed speech. This model provides a useful framework to investigate the interaction of multiple factors in verb learning in a complex environment.
Chapter 6

Conclusions

In this chapter, I review the contributions made by this thesis and consider several interesting directions for future research.

6.1 Summary of contributions

The aim of this thesis was to show that by employing nonparametric Bayesian techniques to probabilistically represent verb argument structure and verb alternation classes, a computational model can acquire robust representations of this knowledge from real-world language use and capture key generalization behaviours shown by children. I argued that the probabilistic representations and inference mechanisms employed by these models provide viable means to explain how children can acquire and generalize highly abstract verb knowledge from usage, in a way that can manage the noise and uncertainty inherent in child-directed speech. I demonstrated this hypothesis by building up a series of Bayesian models that acquire progressively more complex knowledge, and showing how in each case the behaviour of the model reflects known patterns in child language development. These models offer researchers the power to capture developmental phenomena with breadth unavailable in existing computational accounts of verb learning.

Significant theoretical perspectives in language acquisition, including usage-based ap-
proaches (e.g., Langacker, 2000; Goldberg, 2006) as well as earlier views (e.g., Pinker, 1989), have suggested that children rely on strong regularities between form and meaning, particularly with respect to learning verb argument structure. In Chapter 3, we examined the question of how much children may learn about verb meaning by attending to the syntactic aspects of argument structure. Using a Dirichlet process mixture model as a token-level clustering method, we showed that by attending to simple syntactic argument features, a computational model can acquire abstract representations of argument structure that can reasonably distinguish the senses of the highly polysemous verb get. The pattern of acquisition of these senses (as distinguished by the acquired argument structures) largely reflects child behaviour: in most cases, more-frequent senses are easier to learn. However, the model is unable to fully account for the order of acquisition of polysemous senses as observed in children, which provides motivation for the more-powerful models of acquisition described in later chapters. Few computational approaches to verb learning explicitly address the problem of multiple senses of a single verb type, and this model demonstrates the potential value of token-based models of verb learning to investigate polysemy.

In Chapter 4, we showed how verb alternation classes contribute to generalization in verb learning. We developed a hierarchical model that is capable of acquiring knowledge of verb argument structure at multiple levels of inference simultaneously, and demonstrated its behaviour using a large naturalistic corpus of child-directed speech. To our knowledge, this is the first computational model that jointly acquires both verb argument structure and alternation classes from real-world data. Using the dative alternation as a key example, we showed how abstract knowledge of alternations can be generalized to new verbs to support learning. Since the model permits simulations of verb learning in a broad context of many verbs, many argument structures, and many alternations, this simulation of dative acquisition has a much higher degree of realism than in previous work. Moreover, our model has strong implications for developmental theories of lan-
language. By simultaneously learning at multiple levels of abstraction, the model connects a token-level, usage-based representation of language with the highly abstract knowledge necessary to represent and generalize alternations. By representing alternation classes probabilistically, we obtain a kind of weak abstract representation potentially compatible with those suggested by Fisher (2002a), where early abstractions permit some generalizations but not others, and become more adult-like over time.

We augmented this verb class model in Chapter 5 to acquire associations between form and meaning in verb argument structure, and to generalize this knowledge appropriately via the syntactic and semantic aspects of verb alternations. Using a set of general semantic features of verbs, we complemented the syntactic argument features used in the previous chapter to permit detailed investigation of the interactions among multiple factors in verb learning. We captured children’s ability to generalize between form and meaning in two distinct ways. Firstly, we showed that the model can use the alternation pattern of a novel verb to infer aspects of the verb’s meaning, as do children. The model also predicts a novel interaction between the syntactic frame and the animacy of the subject, two possible cues in the alternation that could be used to determine verb meaning. Secondly, we showed that the model can use aspects of the meaning of a novel verb to infer the syntactic frames in which it may participate, again capturing key generalization behaviours of children. We showed an important effect of input frequency on the generalization pattern, an effect that Ambridge et al. (2011) expected but could not observe, likely due to methodological concerns. These simulations show the value of this model as a useful framework to investigate verb learning in a complex environment.

6.2 Future directions

In this section, I discuss further applications of this work to studies in verb learning, and ways that the most advanced model can be further extended and adapted to different
domains.

6.2.1 Other forms of semantic generalization

In Section 5.3.2, we showed that the verb class model is capable of using semantic aspects of novel verbs to predict their argument structure preferences. This is a direct simulation of recent developmental experiments by Ambridge et al. (2011), wherein children used semantic features to distinguish between verbs that are predominantly intransitive and verbs that could alternate between transitive and intransitive forms.

Ambridge and his collaborators have continued to explore children’s use of verb semantics to predict alternations in argument structure, and it would be worthwhile to continue to use our model to investigate such effects. Additional results have shown that children use both verb frequency and semantic factors to retreat from a variety of overgeneralization errors in verb alternations and similar patterns, including datives (*I said her no; Ambridge, Pine, & Rowland, in press:b), locatives (*I poured the cup with water; Ambridge, Pine, Rowland, & Clark, in press), and un-prefixation (*I roll/unroll; close/*unclose; Ambridge, Pine, & Rowland, in press:a).

6.2.2 Acquisition of verb arguments

By jointly acquiring verb argument structure and verb classes over thousands of verbs from a large corpus of child-directed speech, our model offers a broad perspective of verb learning that was previously unavailable in existing computational approaches. However, as in any computational simulation, we must make certain simplifying assumptions that limit the model’s scope. One assumption, in particular, is that children at this developmental stage can already distinguish various syntactic arguments in the input, such as direct objects and clausal complements. However, children do not necessarily acquire these concepts all at the same time, nor all with the same degree of success. One possible extension of this model would be to also learn to identify the syntactic arguments
themselves.

Such an extension would introduce a very large degree of complexity to an already complex model. It is likely, however, that we could reduce the developmental effort by drawing on existing nonparametric Bayesian approaches to the acquisition of syntactic knowledge. It is possible, for example, that we may integrate our existing model with key aspects of an Adaptor Grammar (Johnson et al., 2007) or Fragment Grammar (O’Donnell et al., 2011) model. By doing so, we can investigate interactions between highly abstract verb knowledge and the acquisition of lower-level components of argument structure. Certain kinds of syntactic arguments are likely more difficult to acquire than others, so how does that impact the acquisition of related alternations? Do certain patterns of type and token frequency (Bybee, 1995) in argument structure and alternations make the participating arguments easier to acquire?

6.2.3 Other problems in language acquisition

While in this work we have focused on modelling syntactic and semantic generalizations in verb alternation structure, the verb class model we present offers an inferential structure that can be adapted to the study of other related phenomena in language acquisition. One possible adaptation, still similar to our current verb research, would be to consider cross-linguistic differences in motion verbs denoting manner and path. In another possible application, we might adapt the model’s inference mechanisms to explain the acquisition of regular polysemy in nouns.

Semantic packaging in motion verbs

Different languages have different tendencies of encoding event information in verbs. A native English speaker might say *The girl runs out of the house*, where the verb *run* encodes both the fact of motion and the manner of motion, running. The girl’s path of motion, heading out of the house, is expressed in a separate prepositional phrase.
A native Spanish speaker, on the other hand, might observe the same scene and say *Ella sale de la casa corriendo*, which glosses in English as *She exits the house running*. Here, the verb encodes the path of motion, while the manner is expressed in a modifier.

In general, speakers of so-called *Satellite-framed* languages (like English, Russian, and Chinese) are more likely to use manner-conflating verbs to describe motion events, while speakers of *Verb-framed* languages (like Spanish, Turkish, and Japanese) are more likely to use path-conflating verbs (Talmy, 1985).

English and Spanish speakers differ in the ways they talk about motion events, which results in different generalizations about the meanings of motion verbs. Naigles and Terrazas (1998) show that English speakers tend to expect novel motion verbs to encode manner, while Spanish speakers expect them to encode path. In addition, the sentence frames themselves also affect how speakers interpret novel verbs. A verb phrase including a path prepositional phrase is more likely to induce a manner verb interpretation (*e.g.*, *She’s kradding toward the tree*), while a simple transitive is more likely to suggest a path verb (*e.g.*, *She’s kradding the tree*). Speakers appear to associate syntactic frames with certain semantic packaging patterns. Moreover, there is a strong interaction between a language’s general conflation pattern (manner or path) and the effect of the syntactic frame. Thus, speakers may represent generalizations about motion verbs at two distinct levels of abstraction: at the level of verb argument structure, and at a level encoding general knowledge about different kinds of verbs. Hohenstein et al. (2004) suggest that children may acquire the generalizations that occur at the level of argument structure up to 3-4 years before they learn that, for example, their language has more manner verbs than path verbs.

We can consider how to adapt the verb class model to explain how these two separate factors may be acquired from the distributional properties of the language. The aim is to show how different levels of abstraction in the model capture the different factors involved in generalization. Given an appropriate representation of verb usages that can
express which semantic event components are encoded by various syntactic elements, the argument structure constructions acquired by the model should be able to associate syntactic structures with semantic packaging patterns. Moving up in the model’s hierarchy, the model estimates the distributions of constructions used by each verb, and in doing so should be able to distinguish manner verbs from path verbs in the lexicon. I expect that when the model is trained on English, by clustering similar verbs it will acquire a relatively strong class of manner verbs and a relatively weak class of path verbs. In contrast, training on Spanish should reveal a relatively weak class of manner verbs and a stronger class of path verbs. This behaviour should capture the distinct conflation patterns of each language, and it should be reflected in the model’s generalizations.

**Regular polysemy in nouns**

As with verbs, nouns often exhibit multiple senses, and the lexicon (in English, at least) appears to contain significant regularity in the ways nouns demonstrate polysemy. For example, English shows a very common polysemous pattern between an ANIMAL sense of a word and a MEAT sense. One might say *I chased the chicken around the yard* (the ANIMAL sense), and also *We ate chicken for dinner* (the MEAT sense). Many animal nouns show this pattern (e.g., *goat, lamb, duck*), although there are exceptions (*We ate cow for dinner*). Srinivasan and Snedeker (2011) have shown that children are sensitive to such patterns of regular noun polysemy, and can use them to extend the meanings of novel words. When learning that a *tima* is a new kind of animal, for example, children also expect that *tima* can also refer to the meat of that animal.

We may be able to adapt our verb class model to show how such patterns of regular noun polysemy could be acquired from the input. As we applied the model in this thesis, it acquires classes of verbs that occur with similar argument structure constructions. In Section 5.3.2, we showed that when the model acquires (for example) a novel change-of-state verb, observed only in an intransitive frame, it compares that verb to *other*
change-of-state verbs to infer that the new verb is also likely to participate in a transitive frame. By comparing the novel verb to other similar verbs, we can infer how to extend its usage.

Likewise, we might use a similar inferential mechanism to acquire classes of nouns that occur in a similar range of senses (in a way, a sense alternation—e.g., ANIMAL/MEAT). If the model were to acquire a novel noun, tima, denoting a kind of animal, it may compare that noun against other classes of animal nouns (goat, duck, etc.) and infer that MEAT would be a likely extension of the noun’s meaning.

6.2.4 Probabilistic VerbNet

This work has promising applications beyond modelling child language acquisition. One of the main organizing principles in our model is to form classes of verbs with similar alternation patterns and similar meaning. This, essentially, is the same principle behind VerbNet (Kipper et al., 2008). VerbNet is a large-scale lexical resource, derived from the work of Levin (1993), that encodes the syntactic behaviour and relational semantics of thousands of English verbs. VerbNet has been used in a wide variety of applications, including semantic role labelling (Swier & Stevenson, 2004), the creation of semantic parse trees (Shi & Mihalcea, 2005), and implicit argument resolution (Gerber & Chai, 2010). However, one of the drawbacks of a lexical resource like VerbNet (and other existing predicate lexicons, such as FrameNet; Fillmore, Johnson, & Petruck, 2003) is that its specifications are definitional: VerbNet is a hand-built resource that is restricted to the core usages of verbs in each class. It does not adequately account for natural kinds of flexibility in verb use. People often use verbs in productive ways that go beyond a rigidly defined class structure. In ongoing collaborative work with Martha Palmer and others at the University of Colorado, we have been investigating ways to combine the strengths of VerbNet with the flexible probabilistic approach offered by the verb class model in this thesis.
In the first work resulting from this collaboration, Bonial et al. (2011) take initial steps towards augmenting VerbNet with probabilistic information about coercive constructions. In the form of coercion considered in this work, a verb can be used in “atypical” contexts to extend its meaning beyond its typical usage. For example, the action verb *blink* is typically considered to be strictly intransitive. However, we might use the verb productively to say *She blinked the snow off her lashes*, giving an interpretation whereby the object (*the snow*) is causally affected and changes location. By using the verb productively in a Caused-Motion construction (Goldberg, 2006), the construction coerces the meaning of *blink*.

In this work, we used manually annotated resources to construct a probabilistic profile of the Caused-Motion construction, in order to identify which classes of verbs are likely to participate in the construction as “core,” or definitional, usages, and which are likely to be coercions. We also applied the verb class model from this thesis as an unsupervised method to obtain a similar profile. We found that both methods provide useful and similar information about the coercibility of the Caused-Motion construction. Moreover, the unsupervised method based on our model provides much broader coverage of verbs and verb classes in the lexicon, while avoiding the high cost of manual annotation.
Appendix A

Dirichlet process mixture model

The Dirichlet process mixture model in Chapter 3 is implemented in the OpenBUGS modelling framework (Lunn et al., 2009). For reference, I include the model specification below.

# Bugs model file for learning verb polysemy.

# NOTE ON INDEXING:
# Because of the variable number of dimensions in each feature, we need
# to use an offset index to simulate a ragged array.
# Instead of indexing verb i, feature j, value d by [i,j,d],
# we collapse j and d into one dimension and use the offset value to
# figure out where d starts for each j.
# theta[z[i],offset[j]:offset[j+1]-1] defines the distribution
# for the cluster associated with verb i, feature j, all dimensions.

model {
    for (i in 1:M) {
        for (j in 1:nfeat) {
            for (d in 1:D[i]) {
                theta[z[i], offset[j] + d - 1] ~ ...
            }
        }
    }
}

103
### Appendix A. Dirichlet process mixture model

```r
y[i,j] ~ dcat(theta[z[i],offset[j]: (offset[j+1]-1)])
```

```r
# z is the partition vector for the M usages
z[i] ~ dcat(cprobs[])
```

```r
# Logical array of cluster membership
# C is the max. number of clusters, but it’s a Dirichlet process,
# so make C larger than the expected number
for (k in 1:C) {
  SC[i,k] <- equals(k, z[i])
}
```

```r
# Concentration parameter for the Dirichlet process
# larger values imply more clusters
alpha <- 1
```

```r
# We use the stick-breaking process for the DP prior on the clusters
for (k in 2:C) {
  p[k] <- r[k] * (1 - r[k-1]) * p[k-1] / r[k-1]
}
```

```r
p.sum <- sum(p[])
```

```r
for (k in 1:C) {
  r[k] ~ dbeta(1, alpha)
cprobs[k] <- p[k] / p.sum
}
```
```
for (j in 1:nfeat) {
theta[k, offset[j]: (offset[j+1]-1)] ~ ddirch(ones[1: (offset[j+1]-offset[j])])
}
Appendix B

Verb class model inference

In this section, I describe the Gibbs sampler used to infer posterior distributions of the parameters in the joint verb class model of Chapter 4. For reference, Figure B.1 provides the plate diagram of the hierarchical model. As mentioned in Section 4.2.4, we initialize the model parameters to random values drawn from the prior distributions specified in the model description. We randomly set each of the $z_i$, denoting the clustering of verb usages into constructions, to one of 10 initial constructions. We set each $c_v$ (for grouping of verbs into classes) to one of 10 verb classes. We then iteratively update each of the model parameters by drawing them from posterior distributions conditioned on the input data and the other parameters in the model. Our sampling scheme is based on the HDP implementation of Teh.\footnote{Teh’s implementation is part of the Nonparametric Bayesian Mixture Models package, Release 1, available at \url{http://www.gatsby.ucl.ac.uk/~ywteh/research/software.html}.} In what follows, I describe the process for one iteration of the Gibbs sampler.

We first sample the set of classifications of verbs into verb classes, that is, the indicator variables $c_v$. For each verb in the lexicon, $v$, we treat $v$ as the only verb yet to be classified, and we draw its class $c_v$ from a distribution conditioned on the classifications of all of the other verbs in the lexicon, denoted $c_{\setminus v}$, along with the set of construction assignments $z$ and the hyperparameters $\alpha_0$, $\beta_0$ and $\sigma$. Using Bayes’s rule, and applying conditional
Appendix B. Verb class model inference

Figure B.1: Plate diagram of the verb class model. $I$ represents the total number of input frames, $J$ is the number of features, $Z$ is the total number of constructions, $V$ is the total number of verbs, and $C$ is the total number of verb classes. $Z$ and $C$ are determined by the model as a result of the data, rather than explicitly specified by the modeller. All other parameters are as described in the text.

To estimate the probability under independence assumptions of the model, we estimate this probability as

$$P(c_v|c_{\backslash v}, z, \alpha_0, \beta_0, \sigma) \propto P(c_v|\sigma)P(z_v|z_{\backslash v}, c_v, c_{\backslash v}, \alpha_0, \beta_0)$$  \hspace{1cm} (B.1)$$

where $z_v$ are the current construction assignments for the usages of $v$, and $z_{\backslash v}$ are the current assignments for the usages of all other verbs. The first part of this estimate, $P(c_v|\sigma)$, is the prior probability that class $c_v$ will be used, and it is estimated using a multinomial distribution parameterized on $\sigma$. $\sigma$ includes a non-zero probability of selecting a new, unused verb class for $v$. The second part of this estimate, $P(z_v|z_{\backslash v}, c_v, c_{\backslash v}, \alpha_0, \beta_0)$, is the likelihood that the pattern of construction use for $v$ \textit{i.e.}, $z_v$ could have been generated...
by the overall pattern for the other verbs already in that class. The likelihood of using
construction \( k \), given class \( c_v \), is estimated as follows:

\[
P(k|c_v, c\backslash v, \alpha_0, \beta_0) \propto n_{c_v,k} + \alpha_0 \cdot \beta_0 k
\]  

(B.2)

where \( n_{c_v,k} \) is the number of usages of construction \( k \) over all verbs in class \( c_v \).\(^2\) This class-
level estimate is smoothed by the global construction likelihood \( \beta_0 \) (the \( k \) above indexes
this parameter at construction \( k \)). Thus, the conditional likelihood of all construction
assignments in \( v, z_v \), can be estimated as:

\[
P(z_v|z\backslash v, c_v, c\backslash v, \alpha_0, \beta_0) = \left( \frac{n_v}{n_{vk1}n_{vk2}...n_{vK}} \right) \prod_k P(k|z\backslash v, c_v, c\backslash v, \alpha_0, \beta_0)^{n_{vk}}
\]  

(B.3)

where \( n_v \) is the total number of usages of \( v \) and \( n_{vk} \) is the number of usages of \( v \) currently
assigned to construction \( k \). \( K \) is the current total number of constructions. Thus, the
above estimate for \( P(c_v|c\backslash v, z, \alpha_0, \beta_0, \sigma) \) defines a probability for each verb class in the
model, including a probability of creating a new class for \( v \). We draw the classification
for \( v \) from the resulting multinomial distribution.

Given the classifications of each verb in the lexicon, the next step is to sample the
parameters of the class-level distributions over constructions, \( \phi_c \). For each verb class
\( c \), we draw these parameters from a Dirichlet posterior, given the current construction
counts over all verbs in the class:

\[
\phi_c \sim \text{Dirichlet}(\langle n_{c_v,k1}, n_{c_v,k2}, ..., n_{c_v,K} \rangle + \alpha_0 \cdot \beta_0)
\]  

(B.4)

As described in Section 4.2, the prior on \( \phi_c \) is the global distribution over constructions,
\( \beta_0 \), scaled by the concentration parameter \( \alpha_0 \). \( \beta_0 \) includes a non-zero probability of
selecting a new, unused construction. The angle brackets denote vectors.

We now sample the parameters for the verb-level distributions over constructions, \( \pi_v \).
This is similar to the previous step, although now the posterior draw uses the construction

\(^2\)We use \( k \) to refer to one of \( K \) total constructions in order to simplify indexing in this discussion. \( z_i \)
refers to the specific construction chosen for usage \( i \), and it may range from 1 to \( K \).
Appendix B. Verb class model inference

counts specifically for the verb $v$, and the prior comes from the verb class-level parameters $\phi_c$:

$$
\pi_v \sim \text{Dirichlet}(\langle n_{v,k_1}, n_{v,k_2}, \ldots, n_{v,K} \rangle + \alpha_1 \cdot \phi_c)
$$

(B.5)

Moving down to the level of argument structure constructions, we sample the feature probabilities $\theta_{jk}$ using the observed feature counts and the current assignments of verb usages into constructions. For each construction $k$, let $n_{jk}$ be a vector of the number of occurrences of each value of feature $j$ over all verb usages currently assigned to $k$. We draw the sample of each $\theta_{jk}$ from a Dirichlet posterior as follows:\(^{3}\)

$$
\theta_{jk} \sim \text{Dirichlet}(n_{jk} + 1)
$$

(B.6)

where $1$ is a vector of ones of the same dimension as $n_{jk}$. This effectively adds one to each of the feature counts, and is a standard approach to provide a weak noninformative (i.e., uniform) prior in a Dirichlet-multinomial relationship.

We can now resample the assignment of each verb usage $y_i$ into its construction, $z_i$. The most straightforward way to sample these construction assignments would be to use a similar process to that used to assign verb types to classes above: treat each usage in turn as if it were the last usage to arrive, and sample its construction given the assignments of all the other constructions. Unfortunately, this process causes the sampler to take very small steps through the parameter space, and due to the size of the input and the resulting complexity of the parameter space, such a method is very slow to converge (Teh et al., 2006). Instead, we use a block sampling method,\(^{4}\) which offers substantially faster convergence. In this method, we resample the full set of assignments $z$, all at once, using the construction distributions for each verb, $\pi_v$, and the feature distributions for each construction, $\theta_{jk}$. For each verb usage $y_i$, the probability of the

---

\(^3\)As mentioned in Chapter 4, for binary-valued features, $\theta_{jk}$ parameterizes a binomial distribution, drawn from a beta prior, rather than the multinomial with Dirichlet prior specified here. We discuss the model in terms of the multinomial and Dirichlet as this represents the more general case.

\(^4\)Adapted from the Nonparametric Bayesian Mixture Models package, Release 1, available at \url{http://www.gatsby.ucl.ac.uk/~ywteh/research/software.html}. 
Appendix B. Verb class model inference

construction assignment $z_i$ is estimated as follows:

$$P(z_i|y_i, v, \theta_{z_i}, \pi_v) \propto P(z_i|v, \pi_v)P(y_i|z_i, \theta_{z_i})$$  \hspace{1cm} (B.7)

where $v$ is the verb type observed in $y_i$. $P(z_i|v, \pi_v)$ gives the prior likelihood of observing construction $z_i$ with verb $v$, estimated using the construction distribution $\pi_v$ (which includes a non-zero likelihood of choosing a new, unused construction). $P(y_i|z_i, \theta_{z_i})$ is the likelihood of observing the same set of features of usage $y_i$ in construction $z_i$. We assume that features are conditionally independent given the construction, so we estimate this probability using the product of feature probabilities:

$$P(y_i|z_i, \theta_{z_i}) = \prod_j P(y_{ij}|\theta_{jz_i})$$  \hspace{1cm} (B.8)

Recall that $\beta_0$ defines a global prior distribution over all constructions in the model, and $\sigma$ gives a prior over all verb classes. We resample these parameters from Dirichlet distributions parameterized on the number of items in each of the corresponding clusters—in the case of $\beta_0$, we use the total number of verb usages in each construction, and in the case of $\sigma$, we use the number of verb types in each verb class:

$$\beta_0 \sim \text{Dirichlet}(n_k_1, n_k_2, ..., n_K)$$  \hspace{1cm} (B.9)

$$\sigma \sim \text{Dirichlet}(n_c_1, n_c_2, ..., n_C)$$  \hspace{1cm} (B.10)

Since the clustering of usages into constructions and the clustering of verbs into classes are both nonparametric, we extend the dimensionality of $\beta_0$ and $\sigma$ to include positive probabilities of selecting new constructions and new classes, respectively. To do so, we use the stick-breaking construction of Sethuraman (1994). We draw up to five additional parameters for $\beta_0$ and for $\sigma$, corresponding to new, unused constructions and verb classes. Additional parameters of $\beta_0$ are drawn from a stick-breaking process as follows:

$$\beta'_m \sim \text{Beta}(1, \gamma_0)$$  \hspace{1cm} (B.11)

$$\beta_0 = \beta'_m \prod_{l=1}^{m-1} (1 - \beta'_l)$$  \hspace{1cm} (B.12)
We draw additional values of $\sigma$ similarly:

\[
\sigma'_{p} \sim \text{Beta}(1, \gamma_1) \quad (B.13)
\]

\[
\sigma_p = \sigma'_p \prod_{l=1}^{p-1} (1 - \sigma'_l) \quad (B.14)
\]

Lastly, we resample the hyperparameters $\alpha_0$, $\alpha_1$, $\gamma_0$, and $\gamma_1$. We use the method of Teh et al. (2006), itself adapted from Escobar and West (1995). We refer the reader to these sources for details.

While this Gibbs sampler is straightforward to implement, due to the complexity of the hypothesis space it may show difficulty in converging on a high-probability result. Since each iteration generally introduces small changes to the current hypothesis, the sampler shows rather low mobility through the parameter space—getting to a high-probability region of the space may require moving through a series of lower-probability regions. To improve the sampler’s mobility and thus speed convergence, we use a simulated annealing scheme similar to the method shown to be effective in the word-segmentation model of Goldwater et al. (2009).

Simulated annealing allows a sampler to make more low-probability changes early in the parameter search, in order to let it more quickly explore new regions of the parameter space. To do so, we use a temperature parameter $\tau$, which starts high and gradually reduces to 1. In each iteration, we raise certain probabilities in the model to the power of $1/\tau$ prior to sampling. When $\tau > 1$, this has the effect of flattening the corresponding distribution, making it more uniform, so that the model is more likely to make low-probability changes. The intent, then, is to make the model more likely to ultimately pass through those low-probability regions to reach a high-probability solution. As we gradually reduce the temperature, the samples again become concentrated in high-probability areas.

We apply annealing to two specific probability distributions: the probability of choosing a class for a given verb type, and the probability of choosing a construction for a given
verb usage (equations B.1 and B.7 respectively). In both cases, $1/\tau = (0.1, 0.2, ..., 1.0)$. The appropriate number of iterations to spend at each increment of $1/\tau$ depends on the complexity of the input data. For the Manchester corpus used in Chapter 4, we use the following sequence, where we use each incremental value of $1/\tau$ for the corresponding number of iterations: $(100, 100, 100, 100, 200, 200, 200, 200, 200, 2000)$. These 3,300 iterations correspond to the burn-in period described in the experiment. For the Thomas corpus used in Chapter 5, we use the following sequence: $(50, 50, 50, 50, 200, 200, 200, 200, 200, 2000)$. Again, these 3,050 iterations constitute the burn-in period.
References


linguistic generalizations?: A grammaticality judgment study.


Incorporating coercive constructions into a verb lexicon. In *Proceedings of the ACL Workshop on Relational Models of Semantics*.


References


References


References


