Bayesian Models of Category Acquisition and Meaning Development

Lea Frermann

Doctor of Philosophy

Institute for Language, Cognition and Computation

School of Informatics

University of Edinburgh

2017

Abstract

The ability to organize concepts (e.g., dog, chair) into efficient mental representations, i.e., categories (e.g., animal, furniture) is a fundamental mechanism which allows humans to perceive, organize, and adapt to their world. Much research has been dedicated to the questions of how categories emerge and how they are represented. Experimental evidence suggests that (i) concepts and categories are represented through sets of features (e.g., dogs bark, chairs are made of wood) which are *structured* into different types (e.g, behavior, material); (ii) categories and their featural representations are learnt *jointly* and *incrementally*; and (iii) categories are *dynamic* and their representations adapt to changing environments.

This thesis investigates the mechanisms underlying the incremental and dynamic formation of categories and their featural representations through cognitively motivated Bayesian computational models. Models of category acquisition have been extensively studied in cognitive science and primarily tested on perceptual abstractions or artificial stimuli. In this thesis, we focus on categories acquired from natural language stimuli, using nouns as a stand-in for their reference concepts, and their linguistic contexts as a representation of the concepts' features. The use of text corpora allows us to (i) develop *large-scale* unsupervised models thus simulating human learning, and (ii) model child category acquisition, leveraging the linguistic input available to children in the form of transcribed child-directed language.

In the first part of this thesis we investigate the incremental process of category acquisition. We present a Bayesian model and an incremental learning algorithm which sequentially integrates newly observed data. We evaluate our model output against gold standard categories (elicited experimentally from human participants), and show that high-quality categories are learnt both from child-directed data and from large, thematically unrestricted text corpora. We find that the model performs well even under constrained memory resources, resembling human cognitive limitations. While lists of representative features for categories emerge from this model, they are neither structured nor jointly optimized with the categories.

We address these shortcomings in the second part of the thesis, and present a Bayesian model which jointly learns categories and structured featural representations. We present both batch and incremental learning algorithms, and demonstrate the model's effectiveness on both encyclopedic and child-directed data. We show that high-quality categories and features emerge in the joint learning process, and that the structured features are intuitively interpretable through human plausibility judgment evaluation.

In the third part of the thesis we turn to the dynamic nature of meaning: categories and their featural representations change over time, e.g., children distinguish some types of features (such as size and shade) less clearly than adults, and word meanings adapt to our ever changing environment and its structure. We present a dynamic Bayesian model of meaning change, which infers time-specific concept representations as a set of feature types and their prevalence, and captures their development as a smooth process. We analyze the development of concept representations in their complexity over time from child-directed data, and show that our model captures established patterns of child concept learning. We also apply our model to diachronic change of word meaning, modeling how word senses change internally and in prevalence over centuries.

The contributions of this thesis are threefold. Firstly, we show that a variety of experimental results on the acquisition and representation of categories can be captured with computational models within the framework of Bayesian modeling. Secondly, we show that natural language text is an appropriate source of information for modeling categorization-related phenomena suggesting that the environmental structure that drives category formation is encoded in this data. Thirdly, we show that the experimental findings hold on a larger scale. Our models are trained and tested on a larger set of concepts and categories than is common in behavioral experiments and the categories and featural representations they can learn from linguistic text are in principle unrestricted.

Lay Summary

Humans represent their knowledge about the world around them as categories, which allow them to efficiently learn about, understand, and interact with their surroundings. Concepts (such as objects or actions) are grouped into categories based on common features. For example, sparrows and finches are members of the category bird, because they share features such as their appearance (they have wings and feathers), and their behavior (they can fly, they sing). Established categories can be used to generalize: Observing an unknown creature with feathers that sings and flies allows to infer that it is likely a kind of bird, based on the established knowledge about that category. Categories are fundamental cognitive building blocks: they determine how humans perceive and interact with the world.

Previous research has established three important characteristics of category learning. First, categories are learnt incrementally. Humans observe input over time and integrate the information from those observations into their category knowledge immediately, incrementally improving their representations. Secondly, categories and their associated features are learnt together and mutually improve each other: Knowing that both birds and finches are members of the category bird helps to extract representative features for the category. At the same time, knowing that having feathers is a feature of all members of the category bird helps to categorize unfamiliar objects. Thirdly, categories and their features are flexible and adapt over time. For example, expert education in ornithology establishes increasingly specialized features which may change the representation of members of the bird category.

In this thesis we study the three phenomena described above using techniques from computational modeling. Traditionally, human learning has been investigated in laboratory studies where participants were given a task, for example to learn categories from a small set of artificial objects, and their behavior was carefully analyzed. The tasks and objects involved in laboratory studies are often overly simplistic, and do not capture the complex environment humans are exposed to and learn from. Computational modeling provides an alternative to laboratory studies for investigating cognitive processes. We develop computer programs that simulate the learning process from input data. The programs and the input are systematically manipulated to explore conditions that must be met for successful learning. Moreover, our computational models investigate learning on a large scale and from complex data.

Specifically, our models learn from natural language texts which are available in large quantities and discuss many aspects of concepts. Exposing our models to naturalistic data allows us to simulate learning for a broad range of categories and their representative features. In addition, language plays an important role during category learning in infants. We study child category acquisition by training our computational models on child-directed language (i.e., collections of child-directed language from parent-child interactions).

The first part of this thesis introduces a model which learns categories incrementally, consistently improving the categories while receiving new information over time. We train our model using both general (news) text and child-directed language and show quantitatively and qualitatively that high-quality categories emerge. Our model, however, does not learn representative features together with the categories. We address this shortcoming in the second part of the thesis, where we present a model which learns categories and their features in one process. We evaluate the categories and features emerging from (a) general (encyclopedic) text and (b) child-directed language, and show that meaningful categories and features emerge in both settings. In the final part of the thesis we investigate the development of meaning representations over time. We explore how concept representations develop with increasing age and general knowledge in children. We also investigate how word meaning changes over centuries by applying our model to collections of texts covering multiple centuries.

Acknowledgements

First and foremost, I want to thank my principle supervisor Mirella Lapata for her constant support and enthusiasm. While leaving me the freedom to work on research problems that interest me most, she provided the essential guidance. Her generous feedback has improved the content and presentation of every aspect in this thesis, and shaped my way of thinking about, approaching, and presenting research problems.

I am also grateful my second supervisor, Charles Sutton, for his insightful comments and invaluable feedback on the mathematical aspects of this thesis.

Thanks to my examiners Michael Frank and Chris Lucas for their time to read through this thesis, and for lively discussions and insightful comments during the viva.

I was lucky to be part of ILCC, which provided a lively, fun and interactive environment for research in both NLP and cognition. Thanks to the members of the Prob-Models group for feedback on my PhD project at various stages, and for invaluable advice on presentation skills, and to the ML-for-NLP crew for a relaxed environment for paper discussions.

Thanks to Annie and Dominikus for taking the time to read through draft chapters of this thesis and provide detailed and valuable feedback.

A large number of people I met during my time in Edinburgh made the past years truly enjoyable – both inside and outside the Informatics Forum. Special thanks go to the level-3 lunch group and to Hadi, my longest-lasting office mate, for the reliable daily distractions from work.

Many thanks to my parents and my sister for all their support over the years.

Thank you, Dominikus, for traveling with me through good and stressful times ♡.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Lea Frermann*)

Table of Contents

Chapter 1

Introduction

From day one, infants are exposed to a complex world of objects, humans, and their interactions. They need to acquire an extraordinary amount of knowledge in order to be able to comprehend their environment and react meaningfully to it. Not only in childhood, but throughout their lives, humans continue to experience novel concepts, problems and situations on a daily basis. How do they acquire and represent the knowledge that allows them to understand and interact with the world? *Structured* mental representations, in terms of *categories* have been shown to underlie fundamental cognitive abilities and influence the way humans perceive and react to their environment. This thesis investigates how categories are acquired, how they are internally represented and how these representations change over time.

Categories are fundamental cognitive building blocks allowing humans to organize their knowledge, and make inferences about the world (Rosch, 1978; Medin and Schaffer, 1978; Murphy and Medin, 1985). We investigate the acquisition and representation of superordinate level CATEGORIES, as collections of taxonomically coherent basic level *concepts* (Rosch, 1978).[1] Categories (such as FOOD or ANIMAL) are groups of concepts (such as *apple* or *cat*) which share important properties. Examples of such properties include their appearance (*apples* and *kiwis* contain seeds), their function (*apples* and *kiwis* are edible), or their behavior (*cats* and *dogs* eat and play). Established categories enable *generalization* and *inference*: by extrapolation from established knowledge about FRUIT, one can infer that an unfamiliar object with a sweet smell and seeds inside is likely edible. Table 1.1 summarizes the terminology and

[1] Throughout this thesis we use the term 'category' to refer to superordinate level categories, and 'concept' to refer to basic level categories.

Term	Explanation	Example
concept	living- or non-living thing (basic level category)	*pear*, *cat*, *train*
category	taxonomically coherent set of living- or non-living things (superordinate level category)	FRUIT, ANIMAL, VEHICLE
feature	individual properties of a concept	{eats, is_furry}
feature type	class of properties	`appearance`, `behavior`
stimulus	mention of a concept in a linguistic context	"Cats are furry."

Table 1.1: Overview of the terminology used in the thesis. We denote *concepts* in italics, CATEGORIES as small caps, `feature types` as true type, and {features} as lists.

typography used throughout this thesis.

Understanding the process with which categories and their representations are formed, has been the subject of significant research efforts both from a behavioral and modeling perspective. Prior research on category learning has often involved a small set of artificial stimuli such as binary strings, or purpose-built objects (Anderson, 1991; Bornstein and Mash, 2010), containing a limited number of prominent features. Humans, however, are constantly confronted with myriads of different concepts; just consider the number of objects a child interacts with at home. Furthermore, observations of concepts are often noisy or incomplete. Humans are capable of detecting and filtering noise, distinguishing relevant features from irrelevant ones. Moreover, humans observe concepts in context, and use the context to infer complex and structured featural representations (Murphy and Medin, 1985). Prior research has predominantly focused on adult categorization, assuming that learners have developed categorization mechanisms and a large number of categories have already been learnt. Children, however, learn categories "from scratch", with access neither to prior category knowledge, nor to sophisticated input processing abilities such as language parsing. It is not clear that results from small-scale studies extend to more natural settings where a large number of categories is being acquired and complex representations must be formed in the face of noisy, naturalistic input and under cognitive constraints such as memory limitations.

This thesis investigates the acquisition and representation of categories from naturalistic input at scale. Specifically, we consider three tasks: Firstly, we look at the *in-*

cremental process of category acquisition (Bornstein and Mash, 2010); secondly, we model the *joint* acquisition of categories and *structured* features (Goldstone et al., 2001; Ahn, 1998); thirdly, we investigate how concept representations *develop* over time (Keil, 1987; Schyns et al., 1998; Aitchison, 2001). We approach these questions from a *modeling* perspective, and train and test our models on *linguistic input*. We expose our models to natural language stimuli extracted from large text corpora in an attempt to approximate (a) the number of categories and features acquired in a realistic setting; and (b) the complexity and richness of the learning environment. We investigate category acquisition from large corpora of general text (e.g., encyclopedic text or news). In addition, we expose our models to child-directed language, modeling the specific problem of category and feature acquisition in infants.

Why computational cognitive modeling? Computational models shed light on cognitive processes *at scale*, and make predictions which can be tested empirically. They impose constraints on the cognitive processes under investigation, and learn from exposure to input. The constraints and the input can be manipulated to systematically explore aspects of human cognition in general, and categorization in particular. Specifically, we develop models of category acquisition within the framework of *Bayesian* modeling. Bayesian models provide a mathematically principled way of formalizing constraints and processes, and have been shown to accurately describe a variety of cognitive phenomena (see e.g., Chater et al. 2010). Computational models provide an opportunity to investigate cognition on a larger scale than behavioral experiments. The majority of previous categorization and feature acquisition models either replicate human categorization behavior under simplistic conditions in the laboratory (Anderson, 1991; Sanborn et al., 2006), or are trained on rich and multimodal input, but evaluated on small-scale problems (Yu, 2005; Frank et al., 2009). This thesis capitalizes on the opportunities that computational modeling provides for investigating the *process* of the acquisition and development of a large number of categories and complex features.

Why natural language input? We train and test our models using natural language stimuli, representing observations of concepts as their mentions in text. Figure 1.1 illustrates our representation of concepts. Our choice of representation is motivated by prior research which showed that linguistic input plays an important role during child category acquisition (e.g., Waxman and Markow 1995). Furthermore, much of the structure of our environment is redundantly encoded in language (Riordan and Jones, 2011). Each observation, or stimulus, consists of the mention of a concept

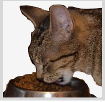
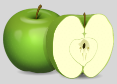
Concept			Natural Language Stimuli	
cat			"*Cats* are furry."	"*Cats* have tails and whiskers."
			"*Cats* are carnivores."	"The *cat* says meow!"
dog			"The *dog* has nice fur."	"*Dogs* have tails."
			"*Dogs* eat meat."	"Look, the *dog* is playing!"
apple			"I'd like to eat an *apple*."	"*Apples* can be red or green."
			"*Apples* grow on trees."	"An *apple* contains seeds"
kiwi			"This *kiwi* is tasty."	"Can you cut me a *kiwi*?"
			"*Kiwis* are green inside."	"*Kiwis* have seeds."

Figure 1.1: Illustration of natural language stimuli provided as input to the models presented in this thesis. Each stimulus contains a mention of a concept (e.g., *cat* or *apple*) in its local linguistic context. Concepts are clustered into categories (e.g., ANIMAL or FRUIT) based on the similarity of the contexts they occur in.

in local context. Linguistic context serves as a representation of the concept's features comprising potentially diverse properties such as perceptual, relational, or other knowledge-based associations. Indeed, it has been shown that child-directed language comprises substantial explicit explanation of non-perceivable features shared among members of categories (Callanan, 1990). The idea that the meaning of a word is characterized by the contexts in which it occurs is well-established as the *distributional hypothesis* (Harris, 1954). In this thesis we extend the distributional hypothesis, assuming that linguistic context is predictive of a word's category. Our models group concepts into categories based on the similarity of the contexts in which they appear.

Classically the term features has been used to refer to lists of necessary and sufficient properties of concepts which have the explicit purpose of defining category membership. Rather than assuming that linguistic context comprises features in the classical sense, we view the features induced by our models as properties which are *associated* with concepts and categories. Concept and category associates have been collected as verbal descriptions of relevant properties of concepts and categories, and have been argued to provide a window into the cognitive representations of concepts in the brain. Typically, such descriptions are collected through feature norming studies (e.g., McRae et al. (2005); Vinson and Vigliocco (2008)) where participants are asked to produce a set of relevant properties of a target concept. Feature norms have been shown to explain a variety of cognitive phenomena and to provide valuable input to computational cognitive models. The work presented in this thesis replaces feature norms with representations derived from concept mentions in contexts in corpora. We assume that words whose referents exhibit differing features are likely to occur in correspondingly different contexts and that these differences in usage can provide an approximation of featural associates. In the context of discussion of the models presented in this thesis, we use the term *feature* interchangeably with *associate*.

The work in this thesis does not address the problem of *word learning* which involves learning a lexicon mapping from words to referent concepts. Instead, we equate words with concepts, assuming that words themselves are instantiations of their referents. In addition, we make the simplifying assumption that concepts and their verbal realization as a written string of letters have a one-to-one correspondence which is known a priori (e.g., the written word ‘dog’ always refers to the concept *dog* and no other word type refers to the same concept). Our models learn to group concepts into categories based on featural commonalities which are acquired through repeated observations of concept instances. In that sense we adopt a cross-situational learning framework (Siskind, 1996).

1.1 Contributions

This thesis investigates a range of category-acquisition phenomena within the unified framework of Bayesian modeling from naturalistic input on a large scale. Our contributions are:

Naturalistic processes. We develop three novel Bayesian models which reproduce four phenomena of human category acquisition and their mental representations which have been established in prior research: Firstly, categories are acquired *incrementally* (Chapter 4). Secondly, categories and their features are learnt *jointly* (Chapter 5). Thirdly, feature representations of categories are *structured* (Chapter 5). Fourthly, featural representations are *dynamic* and flexibly adapt to changes in the learner’s knowledge or environment (Chapter 6). Our models are designed to capture these phenomena in the context of category acquisition in children. They learn categories “from scratch” without initial categorization knowledge and advanced data processing abilities, which young infants do not possess. To the best of our knowledge our work is the first to investigate the joint emergence of features with categories, and their dynamic development over time at scale from naturalistic input. We formalize the above processes within the unified and principled framework of Bayesian modeling. We qualitatively and quantitatively demonstrate the effectiveness of incremental and joint category and feature learning, showing (a) that our models fit incremental human category learning more closely than a previously proposed graph-based model of incremental category learning; and (b) that they learn features which are more interpretable compared to those produced by a knowledge-heavier but cognitively less plausible model of feature extraction from text. Our results provide further evidence to the claim that humans acquire categories by aggregating information over time and by establishing representations which describe their environment increasingly accurately.

Naturalistic input. We show that the above phenomena emerge from our models trained on linguistic stimuli (i.e., mentions of concepts in context; see Figure 1.1). Stimuli are extracted from natural language corpora, which can be noisy or contain irrelevant information. Taken together, the results presented in this thesis reveal that our models are able to detect and represent those aspects which are relevant to concept meaning. They are also able to learn meaningful and richly structured features, which demonstrates that the linguistic stimuli contain rich information about different aspects of properties of concepts. Our models learn categories and their representations from different forms of linguistic input. We learn broad-scale categories from general (news or encyclopedic) text. In addition, we show that our models capture the emergence, representation, and dynamic development of categories and their representations from child-directed language demonstrating that our models capture aspects of category acquisition in infants.

Naturalistic scale. We present our models with large sets of linguistic stimuli in an attempt to approximate the scale and complexity of the environment humans experience. We evaluate the categories inferred by our models against a cognitive gold standard categorization comprising more than 500 target concepts of more than 50 categories. The contexts of our language stimuli are thematically unrestricted in principle, covering a wide range of properties. In this regard, our models learn from rich feature representations approximating the variety of contexts in which concepts are observed in the real world. They model human category acquisition under more realistic conditions than previously proposed models of small-scale category learning from artificial stimuli. We show that Bayesian models of categorization extend to learning problems of naturalistic scale and that Bayesian modeling is a fruitful framework for testing hypotheses about category acquisition, structure, and development.

Relevance for AI and NLP. Beyond the motivation of scientific discovery, understanding human category acquisition may lead to improved mechanisms for artificial intelligence (AI). Humans acquire complex conceptual knowledge, which they then use to understand and reason about the world, highly efficiently and reliably. The ability of machines to represent conceptual knowledge pales in comparison. If human cognition was understood to an extent that allowed implementation in machines, the performance gap could be bridged. The three cognitively motivated computational models introduced in this thesis efficiently learn high-quality categories and their representations, and provide a step towards this goal.

By learning structured knowledge from language input, our models are relevant to the field of natural language processing (NLP), where much research has been dedicated to automatic extraction of information from text. We compare our models against existing models from NLP on tasks including feature extraction from text, and capturing the change of word meaning over centuries. Unlike some of these prior models, our models are knowledge lean (they do not require sophisticated linguistic pre-processing or access to informed prior knowledge), and yet they still perform competitively. The knowledge-lean nature also makes our models straightforwardly applicable to texts across different genres (e.g., text from social media) and languages.

1.2 Thesis Outline

Chapter 2 reviews previous research on category acquisition from an experimental and computational perspective. The first part of the chapter summarizes experimental evidence for strong links between category acquisition and word learning. In the second part of the chapter we position our work with respect to existing computational models of word learning and category acquisition. We discuss existing models in the context of their representation of the learning environment, and the influence it has on the scope of learning problems which can be feasibly modeled.

Chapter 3 introduces Bayesian modeling, the mathematical foundations underlying the models presented in this thesis. We motivate Bayesian modeling as a framework for investigating cognitive phenomena. The second part of the chapter reviews the mathematical paradigm of Bayesian statistics, and the ideas underlying generative Bayesian modeling. We conclude with an overview of Monte Carlo-based sampling methods for approximate Bayesian inference.

Chapter 4 introduces BayesCat, a Bayesian model for large-scale category acquisition. We model category acquisition as an *incremental* process, and investigate the effect of an incremental learning algorithm (by comparison to an ideal batch learner), as well as computational and memory constraints on the learning process and outcome. We study the incremental process of category acquisition by evaluating our model on a large corpus of general texts and on transcribed child-directed speech.

Chapter 5 zooms into specific phenomena of category acquisition: the *joint* emergence of categories and their features in a single process, and the *structured* nature of cognitive representations, as feature types. We present BCF, a Bayesian model which jointly learns categories and their structured representations. We evaluate the quality of the categories and feature representations learnt by our model when exposed to large-scale encyclopedic data. We show that our knowledge-lean, cognitively motivated model performs competitively with a feature extraction model that presupposes a hand-crafted set of rules based on substantial linguistic knowledge. In the second part of the chapter, we investigate the incremental, joint learning process of categories and features in children, applying our model to a corpus of transcribed child-directed speech.

Chapter 6 is concerned with the *dynamic* nature of meaning, and presents SCAN, a dynamic Bayesian model of semantic change. The first part of the chapter investigates how structured concept representations develop and improve over the course of concept acquisition in infants by applying our model to a corpus of transcribed child-directed speech. The second part of the chapter applies SCAN to the task of capturing diachronic meaning change: word meanings change over time and adapt to their speakers' environment. We expose our model to diachronic text corpora and show that it captures a variety of aspects of word meaning change over centuries, and that it performs competitively compared to a range of previously proposed problem-specific models across tasks.

Chapter 7 summarizes our main findings, discusses limitations of our work, and points out directions for future research.

1.3 Published Work

Portions of this thesis have been published previously. The model and experiments presented in Chapter 4 are published in Frermann and Lapata (2015b). A preliminary version of this work was published in Frermann and Lapata (2014). Our work on joint category and feature learning from encyclopedic data (Chapter 5) is published in Frermann and Lapata (2015a). The work on diachronic change of word meaning presented in Chapter 6 is published in Frermann and Lapata (2016).

Chapter 2

Learning Words and Categories

Young children are incredibly efficient learners, and the circumstances and processes underlying the rapid process with which they acquire the skills to interact with and talk about their environment is one of the most widely studied areas in psychology and cognitive science. Our work lies at the intersection of categorization, category- and word learning, as well as the emergence and nature of structured featural representations of categories and concepts. Given the vast amount of prior work in each of these areas, a complete review of prior work is beyond the scope of this thesis.

Infants start learning the meaning of words and the meaning of concepts and categories around the same age, and a broad body of research suggests that the two processes are closely entangled. In the first part of this chapter we provide an overview of these studies and summarize their findings. We then discuss relevant computational models of category and word learning. We review prior work on the featural representations of concepts and categories, their emergence and development in the context of our own models and experiments in Chapters 5 and 6.

Learning categories and learning words is a chicken-and-egg problem: Imagine an infant at the onset of this endeavor hear the word "dog" while observing a situation involving a small furry animal with a tail and two ears that says woof. In principle there are countless potential meanings the child could infer: "dog" might refer to the small furry animal; or to the stretch of land the animal is sitting on; or to its left ear; or to the sound the animal is making; etc. If the child already had acquired conceptual knowledge about the world (which would probably involve a DOG category comprising 'furry living things with tails that say woof', but not a LEFT EAR category), learning the meaning of the word "dog" would be simplified considerably. Conversely, knowing that the word "dog" refers to small furry living things that run and say woof would provide a strong cue for learning the correct conceptual category DOG (rather than a different category, for example one that comprises living things that have particularly prominent left ears).

Extrapolating from this rather contrived example to the complexity of the situations and environments the child is confronted with from day one, the speed and reliability with which children learn to talk and reason about the objects and persons surrounding them seem stunning. The early development of conceptual and linguistic knowledge has been under active research for decades. In the following we review prior behavioral and computational studies which investigate the mutual influence of the two, as well as the learning environment and processes from which they emerge.

2.1 Acquisition of Linguistic and Conceptual Knowledge

In this thesis, we propose models of category acquisition from natural language input, which rest on the assumption that there is a strong relation between linguistic input and emerging categories. Prior work discovered a range of phenomena and biases in early child development which suggest that word learning and learning conceptual categories mutually influence each other (Gopnik and Meltzoff, 1987; Waxman and Markow, 1995; Borovsky and Elman, 2006). The age at which such biases emerge, their specificity to word learning, and the precise mechanisms and direction of influence are subject to considerable debate in the literature, however, their existence has been repeatedly demonstrated in a wide range of studies. Our aim here is to discuss their impact on language and category acquisition.

Linking Words to Objects and Concepts General constraints or biases which guide children at the onset of language acquisition in their hypotheses about potential referents for a novel word have emerged in behavioral experiments over the last decades. Learning a novel word involves (a) mapping the word to a referent (in this review we focus on common nouns referring to objects in the child's environment); and (b) to generalize the meaning beyond the particular situation. For the former challenge children have been shown to assume that unfamiliar words refer to whole objects rather than their parts or properties (*whole-object constraint*, Markman 1991; Hollich et al. 2007), while the generalization problem is influenced by the *taxonomic constraint* (Markman and Hutchinson, 1984; Markman, 1994). The taxonomic constraint refers to the observation that linguistic labels shift children's preference from grouping objects thematically (e.g., cows and milk) towards grouping objects taxonomically, by kind (e.g., cows and pigs). Markman and Hutchinson (1984) presented children with an initial object (e.g., a cow) and two related objects one of which is thematically related (e.g., milk) and the other is taxonomically related (e.g., a pig). They investigated the relations children form in two conditions: in the first condition they provided no label for the initial object ("See this? Can you find another one?"). Children selected the thematically related object (i.e., milk) much more often than the taxonomically related object. In the second condition, the initially provided object was labeled with a novel term ("See this dax? Can you find another dax?"). Children's preferences shifted towards selecting the taxonomically related object (i.e., the pig). This effect has been shown across a variety of studies and paradigms for children as young as 18-months old (Markman, 1991).

The fact that linguistic labels influence the type of inferences children make about relations among objects strongly suggests that language input has an impact on how categories are learnt and represented. Beyond results emerging under laboratory conditions, further evidence for this phenomenon comes from the general patterns of children's linguistic and conceptual development: it has been shown that children's sudden and rapid growth of noun vocabulary (the *naming explosion*) coincides with children starting to sort objects into categories at an age of around 18 months (Gopnik and Meltzoff, 1987). Knowing linguistic labels for objects seems to help infants in constructing and organizing their conceptual representations.

Another constraint on word learning has been put forward which encourages *mutual exclusivity* of labels. Children have a strong preference to associate unfamiliar words with objects for which they do not already know the label (Taylor and Gelman, 1988; Markman, 1994; Xu, 2002). In a typical study, children are presented with two objects, one for which they already know the label (e.g., a doll) and an unfamiliar object (e.g., tongs). When asked to "show the dax" (where "dax" is a novel label) children are more likely to associate the label with the object for which they did not previously know a word (i.e., they select the tongs, Markman and Wachtel 1988.)

The bias towards mutual exclusivity in word learning aligns with a similar tendency concerning concepts and categories: categories are often mutually exclusive (for example, an object cannot be an animal and a fruit). Clearly this is not true in general (for example, an object can be both a fruit and a food). However, children's category representations have been shown to strongly (over-)rely on this assumption of mutual exclusivity. Children have difficulties to acknowledge and learn about inclusive or overlapping categories, for example that an object can be at the same time a doll and a toy (Markman, 1987).

The constraints introduced above provide strong cues for learning the names of objects. But how do children move beyond this task, and learn to name object parts and properties? The mutual exclusivity constraint may be one factor which enables to learn word meanings beyond object labels: If a novel word is used to refer to an object for which the child already knows a label, she may infer that the term refers to one of its parts, its material or another property related to the object instead. Hansen and Markman (2009) show that children learn labels for parts of objects more readily if they already know a label for the object itself, i.e., when mutual exclusivity information is available.

A similar effect has been shown for novel words of different classes, in children who have acquired initial knowledge about linguistic word classes. Children use the linguistic class of a word as a cue regarding potential types of referents. In his pioneering experiments, Brown (1957) presented children with different linguistic forms of a novel nonsense word (*sib*), for example:

(2.1) Do you know what a *sib* is?

(2.2) Have you every seen any *sib*?

(2.3) Show me a picture of *sibbing*.

Children interpreted the novel word as a count noun (2.1), a mass noun (2.2) or an action (2.3), respectively. This result has been replicated widely and has been shown to hold for children as young as two years old (Markman, 1994; Hall et al., 1993), and for words other than nouns: children expect unfamiliar adjectives to relate to properties of objects, or to fine-grained distinctions on the subordinate level (Gelman and Markman, 1985; Waxman, 1990; Waxman and Markov, 1998). Similar effects have been shown for compound nouns (such as grapefruit juice; Gelman et al. 1989). The linguistic form of a novel word raises children's expectations about possible meanings that word might carry.

Words as Invitations to Form Categories[1] Given the general connection between language and concepts, we can now zoom in more closely on the problem addressed in this thesis: The emergence of superordinate[2] level categories such as ANIMAL or FURNITURE as groups of basic level categories (e.g., *dog*, *chair*) from *natural language* input. Basic level categories resemble the perceivable structure of the world. As a consequence, basic level categories (a) tend to refer to concrete objects in the world; (b) are based on salient immediately perceivable features, such as shape, material or color; and (c) are internally homogeneous so that members of the same basic level category share many features while at the same time their features separate them clearly from members of different basic level categories (Rosch et al., 1976). They are cognitively most salient, and are acquired earliest by children (Rosch, 1973, 1978).

While non-linguistic (e.g., visual) cues provide a strong signal for the acquisition of basic level categories, superordinate level categories tend to be abstract groupings which are less obviously coherent. Their meaning is often explained through underlying features which are not immediately noticeable (e.g., ANIMALS breathe, TOOLS have a function). While it seems straightforward to define the basic level category *chair* through a set of observable features, it is difficult come up with a succinct set for the superordinate category FURNITURE. Given this level of abstraction, does language play a central role in superordinate level category acquisition?

A range of studies have investigated the influence of linguistic labels on the acquisition of different levels of categories and reliably found that labels are particularly advantageous (and possibly essential) for the acquisition of superordinate level categories (Waxman and Markow, 1995). Waxman (1990) shows that giving objects noun labels significantly improves preschoolers' ability to categorize objects on the superordinate level. 3-4-year old preschoolers were introduced to 'a very picky doll' who likes only objects of a particular kind. The kinds of objects the doll liked were either of the same superordinate level category (e.g., all animals), or of the same basic level category (e.g., all dogs) or of the same subordinate level category (e.g., all collies). Two conditions were compared: either objects the doll likes are not labeled ("she likes this and this and this"), or the objects are labeled with a 'novel' label (a Japanese

[1] Caption borrowed from Waxman and Markow (1995).

[2] As mentioned earlier, we refer to superordinate level categories as 'categories', and to basic level categories as 'concepts'. For the purposes of this thesis the respective terms are considered synonyms and used interchangeably. We use the traditional terms of basic- and superordinate level categories more heavily in this chapter so that relations to the literature are clear.

noun: “she likes dobits”). Three findings emerged. First, children categorize basic level objects in the unlabeled condition with high accuracy and the availability of a label did not lead to improved performance. Secondly, noun labels had a negative impact on subordinate level categorization performance (however, adjective labels (“she likes dob-ish ones”) were shown to improve subordinate level categorizations (Waxman, 1990, Experiment 3). Finally, children’s superordinate level categorizations improved significantly with available noun labels for objects. In a similar study, Waxman and Markow (1995) showed identical effects in 12-13-month old infants: infants were *only* able to form superordinate categories when a linguistic label was provided, whereas they formed basic level categories irrespective of whether objects were labeled or not.

These results suggest that linguistic cues influence the acquisition of abstract conceptual knowledge which is not immediately reflected in the learner’s perceptual environment. But, does the fact that children are able to group objects into superordinate categories under some conditions mean that they fully conceptualize the underlying meaning of that category? Almost certainly not. Much of this representation is highly structured and dependent on substantial world knowledge (Keil, 1987; Gelman and O’Reilly, 1988). Chapters 5 and 6 address the emergence and development of featural representations of concepts and categories. We review prior research on how this knowledge develops in infants in Sections 5.1.1 and 6.1.1.

In order for these rich representations to emerge, linguistic input likely plays a role that goes beyond providing labels (Gelman and Keil, 1998). While basic level categories are remarkably stable across cultures, sub- and superordinate level categories tend to be culturally informed and are thus more strongly based on conventions rather than on shared perceptual features (see Malt (1995) for a review of both psychological and anthropological perspectives). In order to discover these conventions explicit instruction may be crucial. For example, it seems intuitively plausible to categorize cars based on their color or size, however, cars are conventionally categorized based on features such as their manufacturer, power of their engines, or type of fuel they require. Clearly such categories are difficult to learn from purely perceptual input. Instead, “[w]e need some sort of indication from those who participate in the culture of the things they treat as equivalents and those that are distinguished.” (Brown, 1958, p. 208).

Indeed, adults tend to explicitly explain and point to commonalities of members of superordinate categories (Callanan, 1990). This study found systematic differences in descriptions of basic- and superordinate level categories by adults addressing their 2-

to 4-year old children. While basic level categories were described predominantly in terms of their perceptual properties, parents explicitly pointed out abstract functions and relations pertaining to superordinate level categories. We assume that natural language stimuli used in experiments throughout this thesis encode this kind of input, and consequently usefully approximate the environment of a child learning categories. A second aspect of this learning environment as encoded in linguistic stimuli is the amount of 'training' exemplars children typically encounter. In the labeling studies discussed above, the influence of linguistic labels might have been over-emphasized by the scarcity of training data (although the results suggest that labels allow particularly efficient learning under these circumstances (Waxman and Markow, 1995)).

This thesis investigates the acquisition and development of categories and their representations from natural language input. We assume that this input encodes a substantial amount of the information that is necessary to learn categories. The body of work discussed in this section has shown that conceptual knowledge and language exert mutual influence, and that language input is particularly important for learning higher-level conceptualizations. Constraints on potential word meanings are driven by general conceptual constraints and world knowledge (e.g., the tendency for every object to have only one label parallels the general fact that categories tend to be mutually exclusive). Conversely, the emergence of higher-level conceptual knowledge is driven by linguistic labels and explanations which guide the learner's attention to relevant information, *inviting the learner to form categories* (Waxman and Markow, 1995).

Clearly, however, language is not the only source of information available to young children for the category and word learning process. We now widen our scope and discuss other modalities and cues children exploit for this task. We introduce these modalities in the context of computational models of category- and word learning which leverage different subsets of them. We also discuss the influence of different representations of the learning environment on the scope and kinds of learning phenomena that can be feasibly simulated by computational models.

2.2 Models of Word and Category Learning

Given the substantial body of work on child word and category learning in experimental psychology and cognitive science, it comes as no surprise that a wide variety of cognitively motivated computational models have been proposed over the years which aim to shed light on the process through 'reverse-engineering' the learner. Computational cognitive models implement and simulate a cognitive process, including assumptions about its mechanism and constraints, and allow to systematically examine the effects of different constraints or quantities and characteristics of the input.

We discuss models for both word learning and category learning, given their close relation in both theoretical and computational prior work. Word learning is typically modeled as inferring a lexicon mapping from words to real-world referents and is thus conceptually similar to the problem of inducing a categorization. Here, we review models on a high level, and frame our discussion around the assumptions, motivations and limitations underlying the models proposed in this thesis. Technical details for relevant related models are included in Chapters 4–6 in the context of our own models and experiments. We discuss previous work from two angles: first we look at how the *learning environment* of the human learner is captured in the input representations provided to the computational models, and discuss the impact these representations have on the scale of the learning problem under investigation. Secondly, we discuss the *process* of human learning and ways in which previous computational models have captured its characteristics and constraints.

2.2.1 Input Modalities in Word and Category Learning

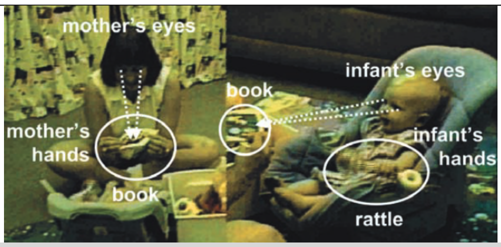
The Multimodal Learning Environment Linguistic input, which was the focus of the previous section, is not the only source of information children receive when they learn words and categories; nor does this learning process happen in isolation: the whole physical environment is rich in cues and highly informative, and together with linguistic and conceptual expertise infants acquire social and motor skills, which enable them to interpret and interact with their environment. In addition, learning is a long-term endeavor: children exploit the fact that words and scenes are repeatedly co-observed over time. Words in child-directed speech refer disproportionally often to objects in the immediate environment. Children use such repeated object-word co-occurrences as cues towards possible referents for labels – a strategy referred to as *cross-situational* learning (Siskind, 1996; Frank et al., 2007; Yu and Smith, 2010; Kachergis et al., 2014).

A variety of models of cross-situational word learning have been proposed (Siskind,

(a) Illustration of the stimuli used in Yu (2005). Image and utterance from Yu et al. (2005).

example	+ "The cow is looking at the little boy."
data	six adults narrating a picture book
input	tuples of { transcribed child-directed utterances features extracted of images from head-mounted camera
test set	lexicon over the set of 12 animals featured in the picture book

(b) Illustration of the stimuli used in Frank et al. (2007, 2009). Image from Frank et al. (2007), utterance added for illustration.

example	 mother's eyes infant's eyes book mother's hands infant's hands book rattle + "Oh, look, a book!"
data	two 10-minute audio and video recordings of mother-child interactions involving a fixed set of toy objects
input	tuples of { transcribed child-directed utterances objects visible to the infant (cf., example) social cues (e.g., gaze of mother and infant; cf., example)
test set	lexicon over 12 toy objects

(c) Illustration of stimuli used in this thesis. Concept mentions are highlighted in italics in the examples.

example	"is there a *train* running on this track", "don't pull the *dog*'s tail", "the ginger pussy *cat*'s called fur ball", "i found the *apple* in the bowl"
data	linguistic mentions of concepts in local context from text corpora
input	transcribed child-directed utterances
test set	$>$ 40 categories comprising $>$ 300 concepts

Figure 2.1: Overview of stimuli used in selected models of word learning from multi-modal data (2.1a, 2.1b), and a comparison to the stimuli used in this thesis (2.1c).

1996; Roy and Pentland, 2002; Yu, 2005; Frank et al., 2007). Yu (2005), for example, propose a model for cross-situational learning of both words and categories from multimodal data, comprising both language and visual input. Their model is based on a three-step process involving (1) recognizing visual features from raw images, (2) clustering those features into visual prototypes, and (3) associating visual prototypes with words from the linguistic input. Input to the model were utterance-image pairs originating from a data set of adults narrating picture books. Individual utterances were paired with raw images taken from a head-mounted camera. The model was evaluated on its ability to induce a lexicon over a set of 12 animals featured in the picture book. The stimuli are illustrated in Figure 2.1a.

Furthermore, children have access to information that goes beyond signal input in various modalities. Even pre-verbal infants are remarkably proficient in interpreting and responding to social cues from the adults they interact with: children interpret intents and follow pragmatic cues (Akhtar and Tomasello, 2000; Csibra and Gergely, 2006). Adults make heavy use of such hints when directing children's attention to objects of interest, i.e., by establishing *joint attention* (Yu and Smith, 2016) and employing different strategies such as prosodic cues, gaze (Yu and Smith, 2007), or actions and gestures like pointing (Yu et al., 2009; Gogate et al., 2000).

These insights have been incorporated in computational models. Yu and Ballard (2007) present a model which learns from audiovisual input stimuli of mother-child interactions which includes both statistical cues from cross-situational occurrences as well as pragmatic cues: each input is labeled with the object of joint attention of mother and child as well as prosodic saliency of words. Their learning algorithm is based on techniques from machine translation, assuming that the child learns a mapping (or translation) from English to an abstract 'meaning language'. They demonstrate the benefit of pragmatic information for a computational word learner.

Methodologically more closely connected to our own approach, Frank et al. (2007, 2009) propose a Bayesian model incorporating social cues in cross-situational word learning. The input to their model consists of transcribed speech from mother-child interactions paired with a representation of the child's visual field in the form of a list of objects present in the environment, as illustrated in Figure 2.1b. Their model jointly learns words (as a dictionary of word-object mappings) as well as to interpret the intended meaning of the speaker (i.e., the object in the environment the speaker is referring to) – without the need to explicitly encode this information in the input as in Yu and Ballard (2007). They show that their joint model not only learns precise lexicons, but also predicts behavioral phenomena such as mutual exclusivity or the use of learnt words for object individuation.

The models discussed so far learn on the basis of a faithful representation of the child's multimodal learning environment, however, the detailed representations come at the cost of a limited scale: all models are evaluated on test sets involving only a handful of target referents and highly restricted input vocabularies. We illustrate the quality and size of the input and evaluation data sets involved in some of the studies discussed above in Figure 2.1. While the quality of the input resembles a child's learning environment, its quantity and complexity does not. Unfortunately, the availability of multimodal data as discussed above is limited and potential further annotations are costly (Frank et al., 2013). It is not clear whether the results still hold with learning problems which are larger in scale, or more complex learning environments such as cluttered scenes. Multimodal corpora available to-date contain short periods of interaction, and consequently do not capture the long-term process of word and category learning.[3]

Learning from Natural Language Input In the experiments in this thesis, we take a step back from a fully, multimodal representation of the learner's environment and instead use large-scale text corpora as input data to our models. Corpora of natural language texts are available in substantial quantities, and the CHILDES database provides a collection of child-directed speech corpora (MacWhinney, 2000). Much of the data consists of transcribed speech resulting from natural interactions of children with their care-takers.

Aside from quantitative motivations, qualitative analyses of linguistic text in general, and child-directed language in particular, revealed that a surprising amount of non-linguistic information is redundantly encoded in language. Riordan and Jones (2011) compare text-based distributional semantics models against models based on human-created feature listings on a semantic clustering task. Feature listings encode diverse information covering multiple modalities (such as visual and functional features, McRae et al. 2005). Purely text-based models performed comparatively

[3] Recent efforts have been made, to create a highly dense longitudinal and multimodal data base of the linguistic development of a single child, by equipping his home with cameras and microphones which constantly monitor his (and his caretakers') development (Roy et al., 2006, 2012). But to this date the data is not publicly available.

to feature-based models, even though they lack the advantage of such rich human-generated knowledge. A comparison of text-based models with models based feature listings encoding sensorimotor knowledge available to children even showed that both systems perform on par when evaluated on child-directed speech data. This may be due to the fact that child-directed speech overwhelmingly addresses the 'here-and-now', i.e., objects, properties or actions of the immediate environment (Veneziano, 2001) and consequently encodes a lot of the information which is also captured in other input modalities (e.g., the visual scene). Fountain and Lapata (2010) show that natural language input is informative for the specific problem of category learning.

Fazly et al. (2010) propose a computational model for cross-situational word learning from large corpora of child-directed utterances, paired with automatically generated semantic descriptions of the scene. These descriptions are sets of abstract symbols corresponding to the spoken words, interspersed with slight disturbances to simulate noise in the learning environment as well as referential uncertainty. Similar to our own child category acquisition experiments, their input is derived from large, longitudinal corpora of mother-child interactions, which allows them to train their model on 20,000 utterances which cover a wide variety of words and objects. They can consequently investigate phenomena such as the developmental process of word learning, or effects of word frequency in the input data.

Category acquisition has also been modeled on a large scale from natural language input. Fountain and Lapata (2011) formalize large-scale category acquisition as an incremental graph clustering problem. They propose an incremental graph-based model of the acquisition of categories comprising more than 500 concepts from both large-scale corpora of generic text as well as child-directed language data (Fountain, 2013). The graph-based model treats the acquisition of concept representations and the clustering of concepts into categories as two separate processes. While our own work leverages a similar representation of the input to the category learner, we formalize the acquisition of categories and their representations as a single process, and within the Bayesian framework. Comparison of our own model with the graph-based model discussed above (Chapter 4) reveals that our Bayesian model qualitatively and quantitatively fits the human category learning process more closely.

In sum, we leverage the support from prior theoretical analyses and computational studies for the fact that natural language input provides a rich environment for both word learning and category acquisition. Using natural language corpora allows us to

Model	Stimuli		Categories
Medin and Schaffer (1978)	6 binary strings	(e.g., 10101, 01000, ...)	2
Anderson (1991)	16 binary strings	(e.g., 0111, 1011, ...)	2
Lee and Navarro (2002)	9 colored shapes	(e.g., 🟥, 🟩, ▲, ...)	2
Bornstein and Mash (2010)	16 physical objects	(e.g., , , ...)	2
This work	$>$ 300 concepts	(e.g., *hat*, *dog*, *car*, ...)	$>$ 40

Figure 2.2: Illustration of stimuli used in selected laboratory studies of human category learning, and the number of stimuli and target categories to be learnt. The bottom line provides our own test set dimensions for comparison.

evaluate our models in a broader setting. Figure 2.1 compares the test set size of previous models of word learning from multimodal input (Figure 2.1a–2.1b) with the test set used in our own studies (Figure 2.1c). We advance previous research by investigating a range of categorization-related phenomena, such as the joint emergence of categories and their features or the dynamic nature of featural representations. All our models are formulated within the Bayesian framework which allows us to express the involved variables and their dependencies in an explicit and mathematically principled manner. We motivate the use of Bayesian methods and related learning paradigms for computational cognitive modeling in the beginning of Chapter 3.

Modeling Human Behavior in the Laboratory Although the process of category and language learning "in the wild" interacts with a myriad of signals from the environment and joint development of other abilities, such as social or sensorimotor skills, much experimental research investigated this process in a laboratory setting. Laboratory studies isolate one phenomenon of interest from as many confounding factors as possible. This creates an 'ideal' data set in the sense that it is free from any confounding factors or idiosyncratic properties of naturalistic data sets. In a typical experiment, participants are taught the category membership, or word meaning, of a set of training stimuli and then asked to generalize to a set of test stimuli. A series of computational models have been developed with the goal to predict the behavioral patterns emerging from laboratory experiments.

Xu and Tenenbaum (2007) pioneered Bayesian modeling of taxonomic word meaning acquisition. In particular, they investigate child and adult patterns of word acquisition for concepts on varying levels of the taxonomic category hierarchy (i.e., subordinate (dalmatian), basic (dog), superordinate (animal)). They show experimentally that humans leverage the statistical structure in the set of (isolated and uncontextualized) examples of word-referent co-occurrences to determine the taxonomic level of a referent of a new word. Their Bayesian model, which incorporates a formalization of this tendency as prior knowledge, replicates a variety of related behavioral phenomena.

A variety of studies of category acquisition in the laboratory and models thereof exist. Anderson's rational model of categorization (Anderson, 1991), for example, was developed to replicate the process with which humans learn categories of abstract stimuli represented by binary features (as illustrated in Figure 2.2) in the laboratory (see also Sanborn et al. (2006)). The model incrementally learns a categorization over those stimuli by integrating new observations into already established categories based on featural similarity. Our model of incremental category learning (Chapter 4) is conceptually similar to the models discussed above, but explores their applicability to a larger number of natural concepts and categories, represented with more complex featural representations.

While laboratory studies allow to observe phenomena of interest free from unwanted confounding factors they also have a range of limitations. Given the temporal restrictions of the experiment, and thus the learning environment that can be simulated, the complexity of training and test scenarios is limited. Stimuli tend to have a small number of manually specified features, and either are concrete objects (e.g., physical objects (Bornstein and Mash, 2010)) or abstract (e.g, binary strings, colored shapes (Medin and Schaffer, 1978; Kruschke, 1993; Lee and Navarro, 2002)). Figure 2.2 illustrates the kind and number of stimuli and categories involved in selected prior laboratory studies of human category learning. The low problem complexity comes with the advantage that it typically allows for exact inference in computational models, avoiding the interference of approximate learning mechanisms. However, it is not clear to what extent the results transfer to more realistic learning problems involving complex concepts and situations. The true environment from which children learn is messy: a multitude of objects, and potential referents are around, and the observational input might be noisy, e.g., visually constrained or highly untypical. Furthermore, in laboratory category learning studies it is difficult to control for the influence of prior knowledge of the participants. Most studies involve adult participants who are equipped with rich knowledge about categorization principles; and even children bring in experience from their interaction with the real world (Neisser, 1987).

In this thesis we develop cognitively motivated Bayesian models of the incremental process of category and feature learning, similar in spirit to Anderson's rational model of categorization, and we apply them to large-scale cognitive learning problems. We show that incremental Bayesian models of category learning explain a range of category acquisition-related phenomena when applied at scale – both in terms of the quantity of available input, its complexity, and the size of our evaluation set.

2.2.2 Learning as an Incremental Process

In the previous sections we reviewed the *what* of computational models of word and category acquisition: the set of phenomena they aim to model and the scope and complexity of the input data. Now, we turn to the *how*, the mechanisms and assumptions, of the learning *process*.[4] There are two prevalent paradigms of (unsupervised) learning: The *batch* paradigm where a learner is presented with a set of training data and systematically extracts information from it, typically holding the data in memory being able to access any data point at any time. The learnt knowledge can be employed after the learning phase is completed. If new data become available, the training phase needs to be re-run from scratch. The *incremental* learning paradigm on the other hand assumes that a learner observes data on-line, over time and integrates extracted information into its state of knowledge immediately. Excessive memory use is not necessary at the cost of always learning from an incomplete data set (due to the ignorance about future observations).

The majority of *machine* learning algorithms adopt the batch learning paradigm, exploiting the availability of excessive memory and processing power. Experimental evidence suggests, however, that (a) human memory is limited and not every observation is stored in memory; and that (b) humans are able to make immediate use of newly encountered information.

Bornstein and Mash (2010) examine object category acquisition (toy objects of two categories differing in color and configuration of their parts) in 5-month old infants. Infants were familiarized with one of the categories in two conditions: one group was exposed to objects of a category for two months in their homes prior to a categorization test, while the second group was exposed to the objects on the day of the test for

[4] These two perspectives roughly correspond to Marr's computational and algorithmic levels of analysis (Marr, 1982).

the first time. During the test, both groups went through a familiarization phase with one category of toys (the same category the home-experience group was already familiar with). Following this, infants were tested with two previously unseen objects, one from the familiar category, and one from the unfamiliar one. Results revealed that infants without prior exposure at home showed signals of learning during the familiarization phase in terms of a change in looking behavior. Their performance in the following categorization task further revealed that they had acquired the category during the familiarization phase. This suggests that their category representations must have formed on-line during the short familiarization phase of just a few minutes. Infants do not require a separate, extended training period in order to be able to make use of inferred category knowledge.[5]

Diaz and Ross (2006) investigate incremental category learning in adults. They show that adults incrementally improve their featural representations of categories, which in turn leads to a better ability to assign objects to categories. Participants immediately use the acquired knowledge throughout the learning process such that their category representations and categorization performance improves over time.

Although the majority of cognitive models of category and word learning rely on traditional batch machine learning techniques,[6] the incremental nature of human learning has been incorporated into models as well. Anderson's rational model of analysis formalizes the incremental categorization process as a non-parametric Bayesian model (Anderson, 1991). The model predicts category membership of observed stimuli based on featural similarity to already established categories. Stimuli are categorized on-line, as they are observed, and once made the categorization decision cannot be revised. Observations are either assigned to an existing category based on feature overlap, or initiate a new category.

While Anderson's incremental learning algorithm involved local approximations, Sanborn et al. (2006) derive an asymptotically exact sequential Monte Carlo algorithm for the same model, in form of a particle filter (Doucet et al., 2001). Without delv-

[5] This is not to say that a prior and extended familiarization phase had no effect: Children who were familiarized with one category in their homes for two months expressed familiarity with the known category from the start of the familiarization phase in the laboratory, and did not show signals of learning.

[6] An important argument in favor of using batch algorithms in cognitive modeling is their representation of an 'ideal learner'. This decision allows the modeler to investigate the learning process under the assumption that the learner has perfect access to and makes perfect inferences based on all available data. Consequently, results can abstract away from inaccuracies in the learning process introduced by temporal or memory limitations of the learner.

ing into the technical details (cf., Section 3.3 for this), particle filters incrementally approximate a target distribution by updating a set of samples from this distribution ('particles') in an on-line fashion with information from new data points, as they become available. Memory constraints can, for example, be modeled by restricting the number of available samples. Particle filters have become a popular learning mechanism for Bayesian cognitive models in recent years and, beyond categorization, have been used to model phenomena like incremental parsing (Levy et al., 2009) or word segmentation (Börschinger and Johnson, 2011, 2012). We develop particle filters to study large-scale human category and feature learning in Chapters 4 and 5.

Beyond modeling human learning in laboratory settings, longitudinal corpora of child-directed speech provide an excellent data source for modeling the long-term incremental process of word learning (Baroni et al., 2007; Fazly et al., 2010) and category acquisition. A graph-based model for incremental category learning from natural language data has been put forward by Fountain and Lapata (2011). The model sequentially observes linguistic stimuli, and constructs meaning representations of concepts from word co-occurrence statistics in the input data. From this representation they infer a categorization of concepts using an incremental graph-clustering algorithm (Biemann, 2006). They incrementally construct a graph where nodes correspond to concepts, and their connection strength is determined by distributional similarity of the linguistic contexts the concepts appear in. Nodes of the graph are clustered into categories based on their connection strength. Constructing semantic concept representations (through co-occurrence statistics) and inducing a categorization of concepts (through graph clustering) are treated as two separate processes. We show that our incremental Bayesian model, while formalizing the process in a unified, and hence cognitively more plausible, process, fits human category learning more closely than the model described above.

2.3 Summary

Learning to represent and to communicate about the rich structure of the environment surrounding us is a fundamental challenge for young infants. In the first part of this chapter we reviewed experimental evidence suggesting that the two tasks of learning language, and learning conceptual knowledge in the form of categories are intertwined and mutually help each other. Building on these findings, in this thesis we model category acquisition from linguistic input.

The second part of this chapter reviewed computational models of word- and category learning. We found that these models are evaluated in limited test settings, comprising the acquisition of a small lexicon or a small number of categories from often artificial stimuli, either due to a sophisticated multimodal representation of the learning environment which is difficult to obtain on a large scale (see Figure 2.1), or due to laboratory experimental settings which are inherently limited in the complexity of the learning task participants are confronted with (see Figure 2.2). In this work we use corpora of natural language to train and test or models on a broader set of categories and features than done previously. The work presented in this thesis is most closely related to the rational models of categorization (Anderson, 1991; Sanborn et al., 2006), and natural language categorization models (Fountain and Lapata, 2011) introduced above. We will discuss these approaches and their relation to our work in Chapter 4 in the context of our own category acquisition experiments.

We now move from the discussion of behavioral experiments and their treatment in computational models, to a technical discussion of the modeling decisions underlying the work in this thesis, and their mathematical foundations. Chapter 3 motivates the use of Bayesian statistics for computational cognitive modeling, and technically introduces its mathematical foundations and algorithms for approximate inference.

Chapter 3

Bayesian Cognitive Modeling and Approximate Inference

This chapter introduces Bayesian modeling, the mathematical framework underlying the models developed throughout this thesis. We begin by motivating Bayesian modeling for cognitive phenomena (Section 3.1), before we move on to a more technical introduction (Section 3.2) and the description of sampling methods for approximate inference (Section 3.3).

3.1 Cognition as Bayesian Inference

Inference is the process of deriving meaningful conclusions from given (possibly uncertain) knowledge. Much of human cognition can be formalized as *inductive inference*, i.e., generalizing from knowns to unknowns, under uncertainty. Humans use established knowledge in order to make inferences about the world, for example when they make decisions (Vul et al., 2014), predictions about everyday phenomena (Griffiths and Tenenbaum, 2006) or learn words (Xu and Tenenbaum, 2007). Categorization is no exception: When learning about a new category, humans need to infer the structure of the category from examples of its members. The knowledge acquired through this process can be ultimately used to make decisions about how to categorize new stimuli. Anderson (1991) pioneered this formalization of the categorization process as inductive inference.

The category inference process described above depends on both established (or prior) knowledge and on observations of stimuli. If observations are scarce or poor in quality, inferences can be based more strongly on prior knowledge. If the prior knowledge is weak or uncertain, the empirical information from the observed data can drive the inferences.

Bayesian statistics mathematically formalize inductive inference. Using the rules of probability, it defines a principled way of drawing conclusions from given information, and provides a means of reasoning about confidence.[1] It assumes that all quantities which are reasoned about are mathematically modeled by random variables. Our goal is to learn the joint probability distribution over all random variables involved. In order to compute this target it is necessary to define a *prior distribution* (encoding existing knowledge) and a *likelihood* (comprising information obtained from observed data). Any inference about the quantities of interest then involves combining the prior knowledge with the likelihood into the *posterior distribution* (encoding our updated knowledge with information from the observed data). Bayes rule formalizes this process and in its intuitive formulation corresponds to,

$$posterior \propto prior \times likelihood. \tag{3.1}$$

The prior, posterior and likelihood are all represented stochastically as (conditional) probability distributions. Slightly more formally, Bayesian inference involves computing conditional probabilities of quantities we want to predict conditioned on quantities that have been observed. We derive Bayes' rule formally in the next section.

The Bayesian inference paradigm has been shown to accurately describe a wide variety of cognitive phenomena (Chater et al., 2010). While few proponents of the Bayesian framework would argue that human cognition actually involves manipulating probability distributions in the brain, the Bayesian paradigm is a useful *descriptive* model of human behavior (corresponding to Marr (1982)'s *computational* level of analysis). An arguably fundamental point of divergence of Bayesian inference from human cognition is the fact that Bayesian inference is *optimal*: a learner using Bayes rule (3.1) to update knowledge will always draw the best possible conclusion from the available data (Jaynes, 2003). However, human behavior is often suboptimal, intuitive and impulsive (Griffiths and Tenenbaum, 2006; Goodman et al., 2008).

[1] Other cognitive modeling paradigms include associative (or connectionist) methods and symbolic architectures. The former paradigm represents knowledge purely in terms of strength of associations and is fraught with difficulties when inferring *structured* knowledge, while the latter is not amenable to graded, probabilistic representation. See for example Tenenbaum et al. (2011) for a more detailed comparison.

Optimal Bayesian inference by exact evaluation of Bayes rule is, however, mathematically complex and computationally expensive. Although equation (3.1) may appear straightforward, the involved probability distributions are often complex and high-dimensional, and the computation of the updates of the prior- to the posterior distribution quickly becomes intractable (Sections 3.2 and 3.3 technically discuss this point). Furthermore, exact Bayesian inference conflicts with human behavior: humans are limited by memory and attention constraints while being able to make inferences within split seconds. Exact Bayesian inference can take hours or days on powerful computers, and often requires vast amounts of memory. Finally, it always leads to the same optimal response under identical conditions, while human behavior exhibits variation (Griffiths and Tenenbaum, 2006; Vul and Pashler, 2008).

A variety of *approximate* inference algorithms have been developed over the past decades which (a) enable tractable Bayesian inference, and (b) better describe human behavior. In this thesis we use sampling-based approximate inference methods, which represent probability distributions as a limited set of realizations of random variables (a sample), with the frequency of realized variable values corresponding to the value's probability under the distribution (cf., Section 3.3).

Various properties make sampling-based approximate inference methods amenable as descriptions of the human inference mechanism. Firstly, they are general methods, in the sense that they can be used to approximate any measures relating to complex functions, and are in no way tied to specific (cognitive) phenomena. Secondly, sampling-based methods can approximate functions of arbitrary complexity which makes them ideal candidates for approximating high-dimensional probability distributions as arising for use in the large-scale models of cognition developed in this thesis. Finally, by varying the size of the sample, sampling-based methods provide an explicit way to approximate the memory-accuracy tradeoff: how many samples are necessary to make inferences with a quality matching human behavior? We investigate this question in the context of large-scale incremental category acquisition in Sections 4.5 and 5.4.

In a machine-learning context it is not uncommon to approximate functions using hundreds or thousands of samples. This seems unrealistic in the face of human processing. Indeed, recent results have shown that in the context of cognitive decision tasks a small set of possibly even a single sample can lead to high-quality predictions (Vul et al., 2014). Results presented in Goodman et al. (2008) for rule-based categorization tasks suggest that individual participants maintain a small sample of rules, leading to individually suboptimal behavior (the aggregate behavior of groups, averaging over individual samples results, however, in optimal responses). Sampling is usually viewed as a useful *description* of the *algorithmic* process underlying human inference (Marr, 1982), rather than assuming that humans physically maintain and manipulate samples of probability distributions in the brain (but see Huang and Rao (2014) for a neural implementation of an incremental sampling algorithm).

The remainder of this chapter provides the mathematical foundation underlying Bayesian statistics and approximate inference. First, we introduce Bayes' rule from a mathematical perspective, and discuss its use within generative models. Next, we introduce sampling-based methods for approximate Bayesian inference (Section 3.3). We discuss Monte Carlo sampling in general (Section 3.3.1), and two specific instantiations: a Gibbs sampler (Section 3.3.2) and a particle filter (Section 3.3.3).

3.2 Bayesian Statistics

Bayesian statistics provides a principled way for reasoning under uncertainty by treating both model parameters and observed data as random variables.[2] This allows to learn distributions over model parameters, which in turn allows to reason about confidence in a particular set of parameters. Probabilities are interpreted as degrees of belief. The ability to reason about uncertainty, or about the degree of belief in a particular model parameterization (often referred to as *hypothesis* in Bayesian modeling terminology), is the fundamental characteristic of Bayesian statistics. Bayesian statistics models the full distribution over parameters. This is in contrast to maximum likelihood estimation which aims to estimate a single best hypothesis, i.e., a point estimate of this distribution. Using the full distribution allows to assess the degree of belief in a particular parameter setting. A point estimate would place 100% of our confidence on one particular parameterization which easily leads to overconfident predictions.

As discussed in the beginning of this chapter, Bayes rule tells us how to combine prior belief with empirically observed evidence (likelihood) into posterior belief,

$$posterior \propto prior \times likelihood.$$

[2] This stands in contrast to the second major statistical paradigm, frequentist statistics. Frequentist statistics treats observed data as random variables, and model parameters as fixed quantities.

More formally, the prior belief encodes the probability of a particular set of parameters (or a hypothesis) θ, before observing any data ($p(\theta)$). The likelihood encodes the probability of the observed data y given this hypothesis $p(y|\theta)$, i.e., how likely is it to observe data y given that hypothesis θ is true? The posterior belief $p(\theta|y)$ corresponds to the probability of a hypothesis given both its prior probability, as well as the likelihood. Updates of the prior beliefs with data-derived likelihoods can be applied repeatedly: the updated, posterior belief serves as the new prior for further reasoning. Bayes' rule is a direct formalization of this process of probability updates:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{3.2}$$

Although this relation is maybe not immediately intuitive, it can be straightforwardly derived from fundamental principles of probability:

$$\begin{aligned} p(\theta|y) &= \frac{p(y,\theta)}{p(y)} && \text{(by definition of conditional probability)} \quad (3.3) \\ &= \frac{p(y|\theta)p(\theta)}{p(y)} && \text{(by the chain rule)} \quad (3.4) \\ &= \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta} && (3.5) \end{aligned}$$

Often we are interested in the relative probability of different hypotheses θ. Since the denominator does not depend on any particular value of θ it can dropped,

$$\pi(\theta|y) \propto p(y|\theta)p(\theta). \tag{3.6}$$

Alternatively, one is often interested in estimating a distribution over parameters θ based on a training set y and use it to make inferences about unseen data $\tilde{y}$. Given that the parameters θ are random variables in our model we do not know their real value, so we have to average (or integrate) over all possible values. This leads to the *posterior predictive* distribution (see Gelman et al. (2014) for an accessible discussion of predictive evaluation methods for Bayesian models),

$$p(\tilde{y}|y) = \int_\theta p(\tilde{y}|\theta)p(\theta|y)d\theta. \tag{3.7}$$

The posterior predictive distribution represents the probability of unseen data $\tilde{y}$ given a hypothesis, weighted by the posterior probability of that hypothesis given the training data y, averaged over all possible hypotheses. Intuitively, rather than attempting to predict the future based on a single hypothesis, predictions are made based on all possible hypotheses, weighted by their probability. This form of reasoning is one of the main benefits of the Bayesian approach.

3.2.1 Bayesian Generative Modeling

Bayesian modeling uses the statistical framework introduced above to fit models to data: given empirical data we want to learn models that explain the data well. Hypotheses correspond to statistical models (or parameterizations thereof) which we can compare and evaluate and our goal is to learn a distribution over models.

Statistical models are systems of probability distributions over sets of random variables. In Bayesian statistical modeling these variables are:

- *Observed variables*, or empirical data y,
- *Hidden variables* θ which conflate all non-observable variables in the model. These include (a) *latent variables* z, which are hidden factors used to explain the structure underlying the observed data; and (b) a vector of *parameters* ϕ, which govern the characteristics of the involved distributions.

The models in this thesis are designed to learn categories from stimuli. Each stimulus consists of a set of words (observed variables). Our models define processes that induce a category structure among the stimuli, by assigning each stimulus a category label (latent variable). Sections 3.2.2 and 3.2.3 discuss distributions we use for modeling the observed and latent variables.

Generative Bayesian models learn a joint probability distribution over observed and hidden variables $p(y,\theta)$. This allows to both infer parameters θ from observations y, but also to inverse the process and, assuming θ is known, generate data y from the model. This generating process is often useful to illustrate the model structure even though practically the model will be used for inference, i.e., learning the parameters given the data.

Taking the Bayesian approach we formulate our models in terms of prior probabilities and likelihoods: we want to learn models that explain the observed data well but we also want to factor in prior knowledge and intuitions we might have about the problems we are tackling. From a modeling perspective, prior knowledge can help to (a) restrict the search space of good hypotheses by directing the model to an a priori likely subspace; and (b) avoid overfitting, i.e., learning parameters which fit the training data too closely, and will generalize poorly towards unseen data.

3.2.2 The Dirichlet-Multinomial Model

In the models developed in this thesis, both observed variables (in our case stimuli consisting of words) and latent variables (such as category labels) are *discrete.* We model observations of discrete variables $\mathbf{x}$ as draws from the Multinomial distribution $Mult(\phi)$ (where $\mathbf{x}$ may refer to either observed or latent variables).[3] Taking the Bayesian approach we draw the parameters ϕ themselves from a prior distribution. We will use the Dirichlet distribution as the prior distribution over Multinomial parameters which itself takes a parameter α (we explain our choice of prior in Section 3.2.2.1). We can summarize our model structure as,

$$\begin{aligned} \mathbf{x} &\sim Multinomial(\phi) \\ \phi &\sim Dirichlet(\alpha), \end{aligned} \tag{3.8}$$

where $\sim$ denotes that the variable on the left is distributed according to the distribution on the right. We now mathematically justify our choice of distributions.

3.2.2.1 Priors and Conjugacy

The fundamental operation in Bayesian inference consists of updating a prior distribution with a likelihood function to form a posterior distribution. Combining arbitrary distributions results in posterior distributions of unknown form which can be difficult to evaluate or sample from. There is a class of well-known priors which, when combined with a likelihood distribution, result in a posterior distribution which belongs to the same class as the prior distribution. A prior with this property is called *conjugate prior* to the respective likelihood distribution.

The Dirichlet distribution is the conjugate prior of the Multinomial distribution. To see this, consider the definition of the Multinomial distribution over a set of c observations $\mathbf{x}$ each of which takes a value $k = 1...K$ under the parameters ϕ,

$$p(\mathbf{x}|\phi, c) = \frac{c!}{\prod_k n_k!} \prod_k \phi_k^{n_k} \qquad\qquad Multinomial(\phi_1 ... \phi_K, c), \tag{3.9}$$

where n_k is the number of observations in $\mathbf{x}$ that take the value k.

[3] It is common in Natural Language Processing to conflate the Categorical and the Multinomial distribution (which generalizes the Categorical distribution to sets of draws) distribution. We will follow the convention of prior work here, and use the term Multinomial throughout even when referring to Categorical distributions, unless otherwise specified.

The Dirichlet distribution is a 'distribution over distributions', i.e., over all possible parameterizations of a K-dimensional Multinomial. A K-dimensional Dirichlet distribution is parameterized through K concentration parameters $\alpha_1 ... \alpha_k$. It is common to set all parameters to the same value, the concentration parameter $\alpha = \alpha_1 = ... \alpha_k$. This results in an uninformative prior, reflecting a priori ignorance about the relative importance of the K outcomes. However, the value of α will support parameterizations of different forms: a small value of $\alpha << 1$ results in Multinomial parameters concentrated on a few outcomes (i.e., a 'peaky' distribution), whereas larger values result in closer to uniform distributions. The probability density function (PDF) of the Dirichlet distribution is:

$$p(\phi|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \phi_k^{\alpha_k - 1} \qquad Dirichlet(\alpha_1 \; ... \; \alpha_K), \tag{3.10}$$

where $\Gamma(\cdot)$ is the Gamma function, a generalization of the factorial to real numbers. Combining the Multinomial distribution from equation (3.9) with the Dirichlet distribution from equation (3.10) leads to another Dirichlet distribution, with updated parameters,

$$\begin{aligned} p(\phi|\mathbf{x};\alpha) &= p(\mathbf{x}|\phi)p(\phi|\alpha) \\ &= \frac{\Gamma(\sum_k n_k + \alpha_k)}{\prod_k \Gamma(n_k + \alpha_k)} \times \left[\prod_k \phi_k^{n_k}\right]\left[\prod_k \phi_k^{\alpha_k - 1}\right] \\ &= \frac{\Gamma(\sum_k n_k + \alpha_k)}{\prod_k \Gamma(n_k + \alpha_k)} \times \prod_k \phi_k^{n_k + \alpha_k - 1} \qquad Dirichlet(n_1 + \alpha_1 \; ... \; n_K + \alpha_K), \end{aligned} \tag{3.11}$$

where the first term in lines 2 and 3 is a normalizing constant.

Note from equation (3.11) that the Dirichlet parameters (α_k) can be interpreted as hypothetical "pseudo-counts" which are added to observations from the data (n_k) and can be interpreted as "imaginary", derived from prior knowledge. The Dirichlet prior thus has a smoothing effect on the data-derived parameters and can help avoid model overfitting.

3.2.2.2 Predicting Observations

We have derived the form of the posterior distribution over Multinomial parameters ϕ under the Dirichlet prior. Often, rather than the distribution over ϕ itself, we are interested in the conditional distribution over values of a new observation x_{t+1} (e.g., the distribution over category labels for an unseen stimulus) *given* the distribution over

over all possible parameters ϕ. The conditional distribution over values for x_{t+1} given all other observations $\mathbf{x}$ can be computed by averaging (or integrating) over all possible values of ϕ. Due to the mathematical advantages implied by conjugate prior-likelihood pairs, this integral can be solved analytically. After some algebraic manipulation the conditional distribution evaluates to a very simple form,

$$\begin{aligned} p(x_{t+1}=k|\mathbf{x},\alpha) &= \int Multinomial(x_{t+1}=k|\phi)\ Dirichlet(\phi|\mathbf{x},\alpha)\ d\phi \\ &= \int \phi_k \frac{\Gamma(\sum_k n_k+\alpha_k)}{\prod_k \Gamma(n_k+\alpha_k)} \prod_k \phi_k^{n_k+\alpha_k-1}\ d\phi \\ &= \frac{n_k+\alpha}{\sum_{k'} n_{k'}+\alpha}. \end{aligned} \tag{3.12}$$

For the interested reader, we derive this result in Appendix A. The probability of observation x_{t+1} taking value k equals the number of times value k was assigned to any other observation in $\mathbf{x}$, normalized by the counts of assignments of any value k' in $\mathbf{x}$. This result allows the derivation of efficient learning algorithms as discussed in Sections 3.3.2 and 3.3.3.

3.2.3 Intrinsic Gaussian Markov Random Fields

While Dirichlet priors are intuitive and computationally advantageous when combined with Multinomial likelihood distributions, the kinds of prior intuitions they can encode are limited. One important limitation is the fact that Dirichlet priors cannot capture dependencies between parameter values. There are however classes of problems which naturally exhibit such structure, for example spatial or temporal variation of a phenomenon of interest. Consider a model for the spread of an epidemic: the severity of infection in any area at any time depends on the level of infection at the area's geographically neighboring areas (due to their proximity and interaction between inhabitants), as well as the level of infection in the area at the previous time (due to epidemics spreading smoothly and continuously). This structure should be captured by a good model.

Chapter 6 of this thesis is concerned with a problem of similar structure, namely modeling of change of meaning over time: we model meaning change as a gradual process which goes hand-in-hand with social, economic and generational change in the population of language users. Like discussed in Section 3.2.2, we still assume multinomial likelihoods. However, we use a different family of priors over these distributions which allow us to capture gradual, or smooth, parameter changes. In particular, we use *intrinsic Gaussian Markov Random Fields* (Rue and Held, 2005; Mimno et al., 2008):

$$\begin{aligned} \mathbf{x} &\sim Multinomial(\phi) \\ \phi &\sim iGMRF(\kappa, \mathcal{Q}). \end{aligned} \tag{3.13}$$

We first define Gaussian Markov Random Fields (GMRFs). Afterwards, we introduce their *intrinsic* version (iGMRFs), as well as properties which make them suitable for capturing smooth parameter dependencies in priors in Bayesian models. Our description is based on various introductions and tutorials on (intrinsic) GMRFs, most notably Rue and Held (2005), Paciorek (2009) and Vivalt (2014). We focus on structures directly relevant to modeling temporal development. However, GMRFs can model a wide variety of structured dependencies between parameters. For a thorough introduction please refer to one of the above references.

GMRFs are undirected graphical models over relations between variables. They are represented by a graph $\mathcal{G} = (E,V)$ consisting of a set of nodes V representing variables, and edges E between pairs of nodes, which indicate a dependency relation. Multivariate normal distributions (MVN) define distributions over n-dimensional random vectors $\phi = [\phi_1 ... \phi_n]$. They are parameterized through a n-dimensional mean vector μ, and a $n \times n$ co-variance matrix which encodes dependencies between variables ϕ_i,

$$\phi \sim \mathcal{N}(\mu, \Sigma). \tag{3.14}$$

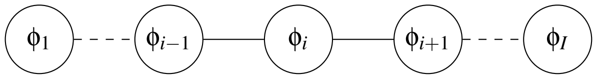
MVNs can be represented graphically through a graph $\mathcal{G}$ as described above: each random variable ϕ_i corresponds to a node in the graph, and edges represent dependencies between ϕ_i. The *inverse* of Σ is the precision matrix $\mathcal{Q} = \Sigma^{-1}$ which explicitly captures the dependency structure between variables ϕ_i, and consequently the graph structure of $\mathcal{G}$. Formally, a random vector $\phi = [\phi_1 ... \phi_n]$ is a GMRF if it follows the distribution

$$p(\phi) = (2\pi)^{-n/2} |Q|^{1/2} exp\big(-\frac{1}{2}\phi^T Q \phi\big), \tag{3.15}$$

by definition of the MVN, and assuming that the mean $\mu = 0$. It is also a GMRF with respect to the graph $\mathcal{G}$ if there is a non-zero entry in Q_{ij} if and only if there exists an edge between nodes i and j in $\mathcal{G}$,

$$Q_{ij} \neq 0 \;\; \Leftrightarrow \;\; \{i,j\} \in \mathcal{G} \;\; \forall i \neq j. \tag{3.16}$$

(a) Graph-representation $\mathcal{G}$ of a first-order GMRF on the line.



(b) The corresponding Precision matrix $Q = \kappa R$ (with $\kappa = 1$).

$$\begin{array}{ccccccc}
1 & -1 & 0 & 0 & 0 & \ddots & 0 \\
-1 & 2 & -1 & 0 & 0 & \ddots & 0 \\
0 & -1 & 2 & -1 & 0 & \ddots & 0 \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\
0 & 0 & 0 & -1 & 2 & -1 & 0 \\
0 & 0 & 0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & 0 & 0 & -1 & 1
\end{array}$$

Figure 3.1: A first-order GMRF on the line with corresponding precision matrix.

For many problems the dependencies between variables are sparse, such that Q has many zero entries, and $\mathcal{G}$ is sparsely connected, which allows for efficient computations. Figure 3.1 displays a graph $\mathcal{G}$ (3.1a) and precision matrix Q (3.1b) representing a structure corresponding to first-order parameter dependencies on the line. Without worrying too much about the values in Q (for now), we can see that that only elements Q_{ij} such that $j - i <= 1$, i.e., immediate neighbors, contain non-zero values, mirroring the graphical dependencies in $\mathcal{G}$. This structure is reminiscent of temporal dependency structure, i.e., each variable in time is connected to (i.e., depends on) its immediate neighbors, and is indeed the structure adopted by the models in this thesis.

The *intrinsic* Gaussian Markov Random Field (iGMRF) is an improper version of the GMRF. Mathematically, this means that the prior is not normalizable (the normalizing constant evaluates to infinity), i.e., it is not a proper probability distribution. While this sounds intuitively unappealing[4] these models have desirable properties which make it a common prior in hierarchical Bayesian models. We discuss the iGMRF of first-order dependencies on the line as shown in Figure 3.1.

The iGMRF of first-order dependencies on the line is defined in terms of independent,

[4] The impropriety of the iGMRF has a lot of theoretical consequences which we will not discuss here, but have been discussed extensively, e.g., in Rue and Held (2005). Most importantly, the iGMRF is usually proper (i.e., normalizable) on a subregion of its probability space, and the posterior distribution arising from combining the iGMRF prior with the likelihood function is also typically proper.

normally distributed, increments between connected variables,

$$\Delta\phi_i \sim \mathcal{N}(0, \kappa^{-1}) \quad i = 1...n-1, \tag{3.17}$$

with $\Delta\phi_i = \phi_{i+1} - \phi_i$. We write the variance in terms of the inverse precision κ^{-1} (just like we used the precision matrix Q instead of the variance-covariance matrix Σ above). The density over ϕ can then be derived, in analogy to (3.14),

$$\begin{aligned} p(\phi|\kappa) &\propto \kappa^{(n-1)/2} exp\big(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(\Delta\phi_i)^2\big) \\ &= \kappa^{(n-1)/2} exp\big(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(\phi_{i+1} - \phi_i)^2\big) \\ &= \kappa^{(n-1)/2} exp\big(\phi^T Q\phi\big), \end{aligned} \tag{3.18}$$

with an appropriately defined scaled precision matrix $Q = \kappa R$ (with R being a matrix capturing the dependencies among the random variables) which turns out to be defined as shown in Figure 3.1b, with

$$Q_{ij} = \kappa \begin{cases} n_i & \text{if } i = j \\ -1 & \text{if } i \sim j \\ 0 & \text{otherwise.} \end{cases} \tag{3.19}$$

Here, n_i refers to the total number of nodes connected to node i, and $i \sim j$ indicates that nodes i and j are connected.

Why is this particular model structure convenient? Conceptually, in priors of Bayesian models which capture the development of some measure over time (or space), it is desirable to be able to only model the development itself without the need to make any claim on the concrete values the measure takes at any time. Often this development is measured as a smooth process over time (or space). Since the iGMRF models the local differences in parameter values (between any individual pair of connected variables) but not the values of those variables, we can achieve exactly this.[5]

[5] This property is mathematically enabled by the impropriety of the iGMRF. The precision matrix Q is not full-rank, which means that there exists at least one vector $\mathbf{y} \neq 0$ such that $Q\mathbf{y} = 0$. For the matrix in 3.1b exactly one such vector exists (Q has *nullity* $k = 1$), namely the vector $\mathbf{y} = \mathbf{1}$, because all rows in Q sum to zero: $\sum_j Q_{ij} = 0 \ \forall i$. Conceptually, these vectors correspond to directions that the iGMRF "has nothing to say about". Mathematically it means that iGMRFs with nullity k are invariant to the addition of a polynomial of degree $< k$. An iGMRF with $k = 1$ is consequently invariant to the addition of polynomials of degree 0, i.e., constants. The degree of a polynomial is the maximum of all exponents of its variables. A polynomial without variables is a constant and necessarily has degree zero. And practically, for the iGMRF on the line with first-order dependencies (Figure 3.1) it means that no global mean of the parameter values exists – the iGMRF has nothing to say about their values, only about their relative difference.

The full conditional distributions of the iGMRF illustrate the model's invariance to the addition of a constant to the global mean,[6]

$$\phi_i|\phi^{(-i)},\kappa \sim \mathcal{N}\Big(\frac{1}{2}(\phi_{i-1}+\phi_{i+1}),\frac{1}{2\kappa}\Big) \quad 1<i<n, \tag{3.20}$$

where $\phi^{(-i)}$ denotes the vector ϕ except element ϕ_i. The value of any variable ϕ_i is normally distributed around the weighted average of the values of its neighboring parameters. The allowed flexibility between values between connected variables (i.e., the "tightness" of the normal distribution) is regulated through the precision parameter κ.

3.3 Bayesian Inference

So far, we have introduced the framework of Bayesian statistical modeling, and discussed the types of distributions which are commonly used as priors or likelihood functions in Bayesian models, and play a central role in the models presented throughout this thesis. While we demonstrated characteristics and limitations of these distributions we did not yet discuss how their parameters can be estimated from data. We now introduce algorithms for Bayesian parameter estimation.

Given a data set y and a model specified in terms of a set of parameters θ, the goal of Bayesian inference is to estimate a distribution over values for θ given the data y. Bayesian inference estimates the full posterior distribution over all possible parameter values (rather than the *best* value) and consequently captures uncertainty about a particular set of values.

Bayes' rule provides a mathematical definition for how to compute this quantity,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)}. \tag{3.21}$$

Practically, however, for all but trivial models exact computations of these quantities is infeasible because the integrals involved become increasingly difficult to compute with a growing parameter space.

Below, we introduce Monte Carlo sampling which provides a way to work with complex probability distributions indirectly, by representing them through a set of samples. Sophisticated versions of Monte Carlo methods have been developed which allow

[6] With 'global mean' we refer to some value μ such that $\mu = E\phi_1 = E\phi_2 = \ldots$.

working with *unnormalized* distributions,

$$\pi(\theta|y) \propto p(y|\theta)p(\theta), \tag{3.22}$$

avoiding the need to compute the integrals involved in the normalizing constant in (3.21). We will provide a brief introduction into approximation through Monte Carlo integration. We then present two instantiations of Monte-Carlo samplers which we will use for parameter inference for our models throughout the thesis:

- A Gibbs sampler, which is a batch inference algorithm that produces parameter samples through repeated iterations over the data, repeatedly updating its parameters according to the unnormalized $\pi(\theta|y)$.
- A Particle filter, which is an incremental inference algorithm and sweeps over the training data only once. It propagates a set of samples (called particles) and immediately updates each sample independently with information extracted from newly encountered data points.

Gibbs samplers are popular inference algorithms which are frequently used for approximating high-dimensional posterior distributions arising in Bayesian models like those discussed in this thesis. The batch procedure underlying the Gibbs sampler, however, seems at odds with characteristics of human learning: humans have memory limitations – they do not memorize large sets of data and perform systematic inference on them. Instead, they use the information of individual observations to make inferences immediately or to update their knowledge (Bornstein and Mash, 2010; Diaz and Ross, 2006). The particle filter resembles this procedure more closely. Chapters 4 and 5 will compare the category acquisition process emerging from our models under the batch Gibbs sampler and the incremental particle filter.

3.3.1 The Monte Carlo Method

The Monte Carlo (MC) method (Hammersley and Handscomb, 1964; MacKay, 2002, Ch., 29) provides a way of approximating complex functions (such as probability distributions) which are impossible or infeasible to evaluate directly. Functions of interest are approximated by *simulation*: a set of samples from the distribution is simulated, and all further computations are carried out on the sample. This works because the expected value of any function, irrespective of its complexity, can be approximated arbitrarily accurately through the mean of independent and identically distributed (iid) samples from the function,

$$\begin{aligned} \mathbb{E}f(\theta) &= \int p(\theta)f(\theta)d\theta && (3.23) \\ &\approx \frac{1}{N}\sum_{i=1}^{N} f(\theta^{i}) \qquad\qquad \theta^{i} \sim p(\theta). && (3.24) \end{aligned}$$

By definition[7] (given in (3.23)) the expected value of a function $f(\theta)$ with respect to random variable θ which is distributed according to $p(\theta)$ is the average of all values θ can take weighted by their probability $p(\theta)$. In line (3.24), the computation of the integral is avoided by instead computing the mean of the function of interest evaluated on a set of N samples $\{\theta^{(i)}\}_{i=1}^{N}$ from the distribution $p(\theta)$. The strong law of large numbers *guarantees* that this mean will converge to the expected value with increasingly many samples. Thus, with $N \rightarrow \infty$ the Monte Carlo approximation of any expectation becomes exact.

We are interested in using the Monte Carlo method for approximating complex probability distributions, i.e., posterior distributions in Bayesian models. How does that relate to the MC definition in terms of expectations above? We can formulate the probability of any value (e.g., $\theta = 5$) in terms of an expectation,

$$p(\theta = 5) = \mathbb{E}\, I_{\{5\}}(\theta). \tag{3.25}$$

where $I_{\{5\}}(\theta)$ is an indicator function which evaluates to 1 if $\theta = 5$ and is 0 otherwise. A Monte Carlo approximation of a probability distribution over all possible values z within the support of $p(\theta)$ can thus be written as,

$$p(\theta = z) \approx \frac{1}{N}\sum_{i=1}^{N} I_{\{z\}}(\theta^{(i)}) \qquad\qquad \theta^{(i)} \sim p(\theta). \tag{3.26}$$

This rather abstract procedure of approximation by simulation is actually very intuitive. In the context of Bayesian inference, simulation refers to generating values from an underlying probability distribution, but it equally works for real physical simulation of events. Imagine, for example, one wants to find out whether a die is fair or not, i.e., whether the distribution $p(\theta)$ over all possible outcomes of die rolls $\theta \in \{1...6\}$ is uniform. This distribution can be approximated by rolling the die N times (i.e., drawing N samples from $p(\theta)$), recording the outcomes, and computing the distribution over

[7] We use $p(\theta)$ as a generic distribution, which may of course depend on further variables, but we drop those here for ease of notation.

outcomes using (3.26). The estimate for $p(\theta)$ is guaranteed to become increasingly accurate with more samples (i.e., rolls of the die).

Unfortunately, plain Monte Carlo simulation as discussed above is often practically infeasible. The complex functions of interest cannot be simulated (or sampled from) efficiently (or at all). A wide range of sophisticated sampling techniques based on the Monte Carlo principle have been developed which avoid the explicit evaluation of the function of interest, and we discuss two of them below: Markov chain Monte Carlo (in the context of Gibbs sampling; Section 3.3.2) and importance sampling (in the context of particle filtering; Section 3.3.3).

3.3.2 Gibbs Sampling

Gibbs samplers provide a way to obtain samples from distributions which can be only evaluated up to a constant, employing the strategy of Markov chain Monte Carlo sampling. We first describe the idea underlying Markov chain Monte Carlo, and then describe the Gibbs sampler.

3.3.2.1 Markov Chain Monte Carlo

Direct iid. sampling from the posterior distribution as required in plain Monte Carlo sampling is often intractable. Rather than generating truly independent and identically distributed samples, it is often more straightforward to draw samples $\{\theta^1,...,\theta^N\}$ which are slightly dependent. Samples can be drawn according to a Markov Chain defined according to $p(\theta)$ (Hastings, 1970). A Markov chain is essentially a random walk over a graph, where vertices (called 'states') correspond to possible values of θ, and the outgoing edges from each vertex define a probability distribution over all next states conditioned on the current state. It satisfies the *Markov* property in the sense that the following state is independent of all previous states given the current state. We can perform a random walk over this graph in steps $n = \{1...N\}$, and draw θ^n conditioned on the previous draw θ^{n-1}:

$$\theta^n \sim p(\theta^n|\theta^{n-1}). \tag{3.27}$$

Note that (a) we generate samples through repeated evaluation of *local* probability distributions, or state transitions, and thus avoid to evaluate the complex distribution

$p(\theta)$ (which prevented us from using plain Monte Carlo simulation); and (b) that the draws from $p(\theta)$ are no longer independent and identically distributed so that the strong law of large numbers used in the motivation of Monte Carlo methods no longer holds. Under some circumstances the sequence of states $\{\theta^1, ..., \theta^n\}$ visited in the random walk corresponds to a sample from $p(\theta)$, i.e., each state is visited with a probability proportional to $p(\theta)$, which means that $p(\theta)$ is the *stationary distribution* of the chain. Conceptually, these conditions include:

- The random walk must be initialized in some way, but the sample resulting from a (long enough) random walk should be independent of the starting point. More formally, after an initial period, the probability of reaching any state θ does not depend on the initial state θ^0.

- In order for a state sequence of a long random walk to be a valid sample from $p(\theta)$ we must make sure that we can in principle visit all areas under the support of $p(\theta)$ at any time during the walk, i.e., we do not want to "get stuck" in a particular sub-space of the distribution. Consequently, our random graph must be highly connected, and guarantee for an infinitely long walk started at any particular state z that the probability to re-visit z in the future is 1.

Concretely a valid Markov chain must be *ergodic*, which means that it must be *aperiodic*, *irreducible* and *positive recurrent*. However, we will leave our introduction on this conceptual level, and invite the interested reader to learn more about these concepts in excellent mathematically rigorous introductions to Markov chains and MCMC such as Bishop (2006) and Murphy (2012).

We will now introduce the Gibbs sampler which is one method for constructing a valid Markov chain for sampling from a target distribution $p(\theta)$.

3.3.2.2 The Gibbs Sampler

The Gibbs sampler (Geman and Geman, 1984; Bishop, 2006) is a Markov Chain Monte Carlo method, which is particularly suitable for sampling from probability distributions over high-dimensional parameters $\theta = \{\theta_1, ..., \theta_I\}$ (i.e., when I is large), as it is the case in the models developed in this thesis. We focus on Gibbs sampling for sampling from the posterior distribution over parameters given data $p(\theta|y)$. The Gibbs sampler constructs an ergodic Markov chain over parameter samples from $p(\theta|y)$ as a sequence

Algorithm 1 The Gibbs Sampler.

```
Initialize the sampler to a random starting point $\theta^0 \leftarrow \{\theta^0_1, \theta^0_2, ..., \theta^0_I\}$
repeat
    Run the sampler for $b$ burn-in iterations
    for each iteration $n = [b+1...]$ do
        for each dimension $i = [1..I]$ do
            $\theta^n_i \sim p(\theta^n_i|\theta_{-i}) = p(\theta^n_i|\theta^n_1, ...\theta^n_{i-1}, \theta^{n-1}_{i+1}, \theta^{n-1}_I)$
        if lag $> \ell$ then
            return a sample from the joint posterior distribution
                  $\theta^n = \{\theta^n_1, \theta^n_2, ..., \theta^n_I\}$
until the desired number of samples has been collected.
```

of samples from full conditional distributions of each individual parameter θ_i. The full conditional distribution for parameter θ_i defines the distribution over values for this parameter conditioned on the current values of all parameters other than θ_i. We denote this set as θ_{-i}:

$$\theta_i \sim p(\theta_i|\theta_{-i}, y) = p(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_I, y). \tag{3.28}$$

The full conditional distributions must be normalized, because it must be possible to draw samples from them. Luckily they are one-dimensional by definition, which typically allows for proper normalization. In our case these full conditionals are discrete distributions over a finite probability space, so normalization is feasible.

Why is this sequence of full conditional distributions a valid approximation of the target posterior distribution, i.e., the joint distribution over $\theta = \{\theta_1, ..., \theta_I\}$? It turns out that the full conditional distribution of any parameter θ_i is proportional to the joint distribution over parameters θ:

$$p(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ...\theta_I) = \frac{p(\theta_1, ..., \theta_I)}{p(\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ...\theta_I)} \propto p(\theta_1, ..., \theta_I). \tag{3.29}$$

The complete algorithm of the Gibbs sampler is displayed in Algorithm 1. The sampler starts with a randomly initialized parameter vector θ^0. It then repeatedly iterates over the components of θ and resamples each θ_i individually from its full conditional distribution (equation (3.28)). Periodically, the current value of θ is returned as a sample. The algorithm terminates when the required number of samples are obtained. There are a few practical intricacies which arise with MCMC samplers in general and the Gibbs sampler in particular:

- The values of parameters θ are usually randomly initialized, which means that the sampler starts off at an arbitrary position in the state space, probably some distance away from the high-probability region under the posterior distribution $p(\theta|y)$. Although it is guaranteed through the ergodicity property that the sampler will ultimately produce samples distributed according to the posterior distribution, it needs a number of iterations to reach this distribution. This initial period is called *burn-in* period. It is difficult to exactly determine the point at which the stationary distribution is reached, and it is common practice to discard a safely large set of initial samples.

- Recall that MCMC samples are generated from a Markov chain producing samples from $p(\theta^n|\theta^{n-1}, y)$ and hence locally correlated. To obtain a set of samples which are as close to iid as possible in an efficient way[8] it is common practice to include a *lag* ℓ and only collect every ℓ^{th} sample (a process called "thinning"). Again setting ℓ to an appropriate value is more of an art than a science.

3.3.2.3 Collapsed Gibbs Sampling for Dirichlet-Multinomial Models

How does this relate to the models presented in this thesis? Recall that we sample values from the posterior distribution over parameters θ. Recall also, that parameters in Bayesian models comprise both variables (e.g., the category of an observation), as well as the parameters governing the distributions from which latent- and observed variables are generated (ϕ).

As discussed in Section 3.2.2.1 we are often not interested in the distribution-governing parameters ϕ themselves[9], but rather in the distribution over value assignments to latent variables of observations $\mathbf{x}$ (e.g., in assigning a category label $i \in \{1...I\}$ to an observation x^j). We showed in Section 3.2.2.1 (equation (3.12), page 37) that the conjugate pair of the Dirichlet and the Multinomial distribution allows to analytically integrate over parameters ϕ.

In the context of a Gibbs sampler, analytically integrating (or *collapsing* or *marginalizing*) Multinomial parameters means that we do not need to resample their value explicitly, but that they are implicitly represented through the *sufficient statistics* of value

[8] This method can be useful to reduce computational and/or memory requirements of computing estimates, but it will not improve the accuracy of the estimates (Link and Eaton, 2012).

[9] Although they can be recovered given an estimated model.

assignments to variables. Collapsing parameters of a model constrains the state space and often leads to more efficient samplers.

Collapsed Gibbs sampling then corresponds to repeatedly sampling each individual latent variable from its full conditional distribution, i.e., the distribution over values assigned to variable x^j conditioned on the values assigned to all other variables $\mathbf{x}^{-j}$ in the model, while implicitly marginalizing over the data-generating distribution ϕ. These full conditional distributions evaluate to a very simple form:

$$\begin{aligned} p(x^j = i|\mathbf{x}^{-j}, \alpha) &= \int p(x^j = i|\phi)p(\phi|\mathbf{x}^{-j}, \alpha)d\phi && (3.30) \\ &\propto \frac{n_i^{-j} + \alpha}{\sum_{i'} n_{i'}^{-j} + \alpha}, && (3.31) \end{aligned}$$

where n_i^{-j} is the count of assignments of value i to any observation, excluding counts related to observation x^j. The probability that observation j has value i is proportional to the number of times value i is assigned to any other observation $\mathbf{x}^{-j}$. In this way values are repeatedly re-assigned to variables without ever explicitly representing the parameter ϕ in the sampler. See Appendix A for a detailed derivation.

3.3.3 Particle Filtering

Markov chain Monte Carlo methods, like the Gibbs sampler introduced above, iterate repeatedly over the entire input data set in order to produce samples from the posterior distribution. This can be undesirable for various reasons. The available data might grow over time and updating the posterior estimate requires to re-run the sampler, which can be expensive. Furthermore, a (vanilla) MCMC sampler holds the entire data set in memory, which is implausible from a cognitive point of view. For learning to occur it is not necessary to have access to all available information or hold it in memory. In this section we introduce particle filters, a method for *incrementally* estimating a posterior distribution.

Chapters 4 and 5 in this thesis introduce Bayesian models for investigating the incremental process of human category learning, where novel information from observed stimuli is immediately used to update the category representation (Bornstein and Mash, 2010; Diaz and Ross, 2006). Particle filters provide a mathematically principled way to model incremental learning of Bayesian model parameters (Sanborn et al., 2006).

Particle filters estimate the posterior distribution over unobserved parameters (e.g., possible categorizations of stimuli) $p(\theta|y)$ in real time, as data is observed. Each time point t corresponds to an observation of a data point (e.g., stimulus) y_t. We use θ_t to denote a concrete parameterization at time t (e.g., a specific categorization of all stimuli $\mathbf{y}_{1:t}$ observed up to time t). At each time t, we want to estimate the posterior distribution over parameters given all data observed up to that time $p(\theta_t|\mathbf{y}_{1:t})$. Particle filters maintain an approximation of these distributions as a set of weighted samples:

$$p(\theta_t|\mathbf{y}_{1:t}) \sim \left\{ \left(\theta_t^{(i)}, w_t^{(i)} \right) \right\}_{i=1}^{N}, \tag{3.32}$$

where (θ, w) refers to a (sample, weight) tuple, $\{\cdot\}_1^N$ denotes a set of N such tuples, and $\sim$ (by slight abuse of notation) means "is represented as". This set of particles is updated incrementally from a representation at time $t-1$ to a representation at time t, with every incoming stimulus y_t,

$$p(\theta_{t-1}|\mathbf{y}_{1:t-1}) \sim \left\{ \left(\theta_{t-1}^{(i)}, w_{t-1}^{(i)} \right) \right\}_{i=1}^{N} \xrightarrow{y_t} p(\theta_t|\mathbf{y}_{1:t}) \sim \left\{ \left(\theta_t^{(i)}, w_t^{(i)} \right) \right\}_{i=1}^{N}. \tag{3.33}$$

Particle filters use *sequential importance sampling* (SIS) for efficiently and repeatedly computing this update. SIS approximates Bayesian optimal filtering which defines the exact way for recursively estimating a Bayesian posterior distribution but is computationally infeasible (see e.g., (Särkkä, 2013)). A well-known property of SIS is that the approximation of the target distribution decreases in quality over time. In order to alleviate this problem, particle filters involve an additional *resampling* step. Resampling provides a way to periodically re-position the filter to high-probability areas of the sample space.

Figure 3.2 illustrates the particle filtering process, and Algorithm 2 displays it algorithmically. We derive importance sampling (Section 3.3.3.1) and sequential importance sampling (Section 3.3.3.2), before we discuss resampling (Section 3.3.3.3). Section 3.3.3.4 concludes with a brief discussion of Rao-Blackwellized particle filtering.

3.3.3.1 Importance Sampling

At each time t the particle filter maintains a sample from the posterior distribution $p(\theta_t|\mathbf{y}_{1:t})$. As discussed previously, exact sampling from Bayesian posterior distributions is often impossible. Instead, particle filters use importance sampling (IS; Geweke 1989; Bishop 2006) which approximates a complex target distribution $p(\theta|\mathbf{y})$ with samples from a simpler importance distribution $q(\theta|\mathbf{y})$. It uses the following identity of the Monte Carlo principle,

$$\int f(\theta)p(\theta|\mathbf{y})d\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta^{(i)}) \qquad \theta^{(i)} \sim p(\theta|\mathbf{y}) \tag{3.34}$$

$$= \int f(\theta)\frac{p(\theta|\mathbf{y})}{q(\theta|\mathbf{y})}q(\theta|\mathbf{y}) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{p(\theta^{(i)})|\mathbf{y})}{q(\theta^{(i)}|\mathbf{y})} f(\theta^{(i)}) \qquad \theta^{(i)} \sim q(\theta|\mathbf{y}) \tag{3.35}$$

$$= \sum_{i=1}^{N} w^{(i)} f(\theta^{(i)}). \tag{3.36}$$

Equation (3.34) repeats the Monte Carlo principle as introduced in Section 3.3.1. In (3.35) we multiply and divide by the same factor, thus not changing the equation, but reformulating it such that the sampling distribution is now $q(\theta|\mathbf{y})$. In the last step we rewrite the approximation (3.35) in a way that introduces importance weights $w^{(i)}$. Importance weights $w^{(i)}$ correct for the discrepancy between the importance and the target distribution. In order make (3.36) a valid approximation, the importance weights must be defined as,

$$\tilde{w}^{(i)} = \frac{p(\mathbf{y}|\theta^{(i)})p(\theta^{(i)})}{q(\theta^{(i)}|\mathbf{y})}, \tag{3.37}$$

and subsequently normalized such that they sum to one:

$$w^{(i)} = \frac{\tilde{w}^{(i)}}{\sum_j \tilde{w}^{(j)}}. \tag{3.38}$$

And as a result, it is no longer necessary evaluate or sample from the complex target distribution $p(\theta|\mathbf{y})$.

In sum, importance sampling consists of three steps: (1) draw N samples $\{\theta^{(i)}\}_{i=1}^{N}$ from the importance distribution $q(\theta|\mathbf{y})$; (2) compute the unnormalized importance weights (equation (3.37)); (3) normalize the importance weights to sum to one (equation (3.38)).

3.3.3.2 Sequential Importance Sampling

Particle filtering uses the importance sampling procedure described above repeatedly, for obtaining an estimate of the target distribution at each time t. Instead of generating a new sample from scratch at each time, the existing sample from time $t-1$ is recursively *updated* with new information. *Sequential* importance sampling (Doucet et al., 2001) defines an efficient way for recursively updating samples and their associated weights.

The process underlying sequential importance sampling is illustrated in Figure 3.2 (top). Parameters develop through a first-order Markov Process, i.e., at any time the distribution over θ depends only on θ_{t-1}, and observations y_t are independent given parameters θ_t. The posterior distribution over parameters $p(\theta_{1:t}|\mathbf{y}_{1:t})$ can be defined recursively, by updating an existing distribution over parameters $p(\theta_{1:t-1}|\mathbf{y}_{1:t-1})$,

$$p(\theta_{1:t}|\mathbf{y}_{1:t}) \tag{3.39}$$

$$\propto p(\theta_1)p(y_1|\theta_1)\prod_{n=2}^{t} p(\theta_n|\theta_{n-1})p(y_n|\theta_n) \tag{3.40}$$

$$= \underbrace{p(\theta_1)p(y_1|\theta_1)\left[\prod_{n=2}^{t-1} p(\theta_n|\theta_{n-1})p(y_n|\theta_n)\right]}_{p(\theta_{1:t-1}|\mathbf{y}_{1:t-1})} p(\theta_t|\theta_{t-1})p(y_t|\theta_t) \tag{3.41}$$

$$= p(\theta_{1:t-1}|\mathbf{y}_{1:t-1})p(\theta_t|\theta_{t-1})p(y_t|\theta_t). \tag{3.42}$$

where we use the Markov properties introduced above in (3.40). We get (3.41) by separating out the last observation t. We then re-collapse times $t = [1...t-1]$ obtaining one factor corresponding to the posterior distribution at time $t-1$, which is updated with the new information from time t (equation (3.42)).

In order to be able to sequentially estimate this target, we need a recursive definition of the importance distribution,

$$q(\theta_{1:t}|\mathbf{y}_{1:t}) = q(\theta_{1:t-1}|\mathbf{y}_{1:t-1})q(\theta_t|\theta_{1:t-1},\mathbf{y}_{1:t}). \tag{3.43}$$

Following the idea of importance sampling (Section 3.3.3.1) and using (3.42) as the target distribution and (3.43) as the importance distribution, we can define importance weights. It turns out that the importance weights can also be defined recursively. We obtain the weight of the i^{th} particle at time t, $w_t^{(i)}$, by updating its previous weight at time $t-1$, $w_{t-1}^{(i)}$,

$$\begin{aligned} w_t^{(i)} &\propto \underbrace{\frac{p(\theta_{1:t-1}^{(i)}|\mathbf{y}_{1:t-1})}{q(\theta_{1:t-1}^{(i)}|\mathbf{y}_{1:t-1})}}_{w_{t-1}^{(i)}} \frac{p(\theta_t^{(i)}|\theta_{t-1}^{(i)})p(y_t|\theta_t^{(i)})}{q(\theta_t^{(i)}|\theta_{1:t-1}^{(i)},\mathbf{y}_{1:t})} \\ &= w_{t-1}^{(i)} \frac{p(\theta_t^{(i)}|\theta_{t-1}^{(i)})p(y_t|\theta_t^{(i)})}{q(\theta_t^{(i)}|\theta_{1:t-1}^{(i)},\mathbf{y}_{1:t})}. \end{aligned} \tag{3.44}$$

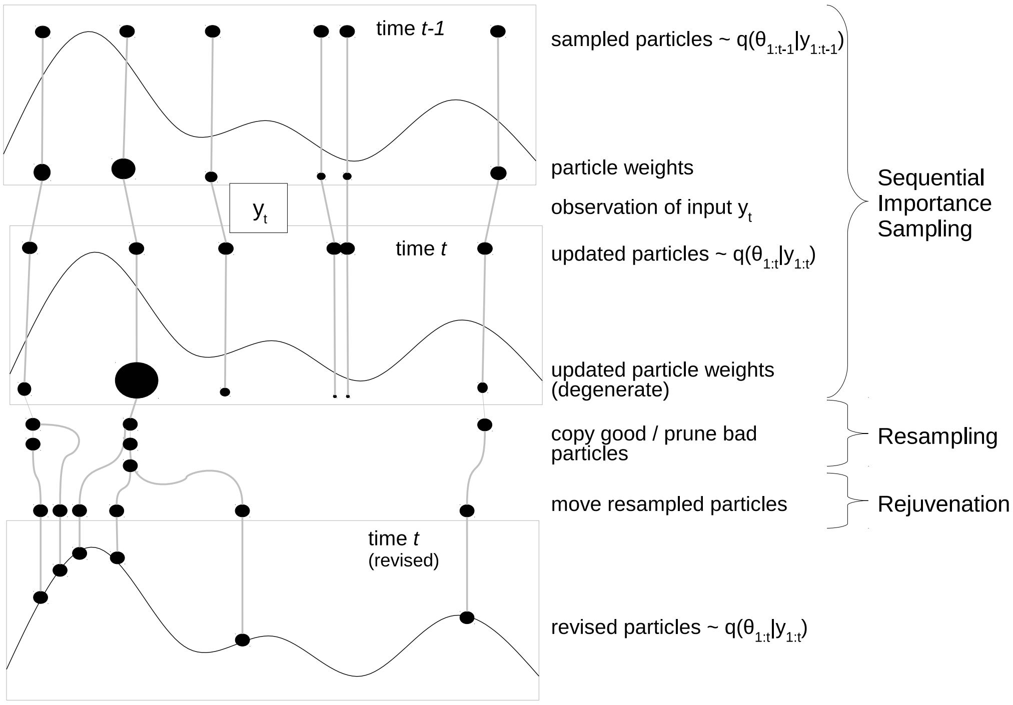

time t-1
sampled particles ~ q(θ1:t-1|y1:t-1)
particle weights
yt
observation of input yt
time t
updated particles ~ q(θ1:t|y1:t)
updated particle weights (degenerate)
Sequential Importance Sampling
copy good / prune bad particles
Resampling
move resampled particles
Rejuvenation
time t (revised)
revised particles ~ q(θ1:t|y1:t)

Figure 3.2: Schematic illustration of the particle filtering algorithm. A set of $N = 6$ particles incrementally approximates the target distribution (graph). Particles are denoted as circles, and circle size represents the particle weights. At each time t, the existing set of particles from $t-1$ is updated with new incoming information, and the particle weights are updated (top; Sequential Importance Sampling). Periodically 'bad' (low weight) particles are replaced with 'good' (high weight) particles (resampling; middle). After resampling, resampled particles are slightly perturbed to increase sample diversity (rejuvenation; bottom).

The Optimal Importance Distribution. The importance distribution $q(\theta_{1:t}|\mathbf{y}_{1:t})$ can be chosen at liberty, and is usually defined such that it is easy to sample from. The (sequential) importance sampler is most efficient, however, when the importance distribution is as similar to the target distribution as possible. The 'optimal' importance distribution is defined such that it minimizes the variance among importance weights (Zaritskii et al., 1976), and is given by,

$$q(\theta_{1:t}|\theta_{1:t-1}, y_{1:t}) = p(\theta_t|\theta_{t-1}, y_t). \tag{3.45}$$

Note that this distribution is *locally* optimal because it is conditioned on the fact that the sequence of sampled parameters remains unchanged. Otherwise the algorithm would cease to be sequential. It corresponds to the posterior distribution over parameters θ_t considering both prior information from the parameter estimate θ_{t-1} as well as information from the current observation y_t. Under the optimal importance distribution, the sample weights correspond to the predictive likelihood of observation y_t:

$$\begin{aligned}
w_t^{(i)} &= w_{t-1}^{(i)} \frac{p(\theta_t^{(i)}|\theta_{t-1}^{(i)})p(y_t|\theta_t^{(i)})}{q(\theta_t^{(i)}|\theta_{1:t-1}^{(i)}, \mathbf{y}_{1:t})} && \text{copy (3.44)} \\
&= w_{t-1}^{(i)} \frac{p(\theta_t^{(i)}|\theta_{t-1}^{(i)})p(y_t|\theta_t^{(i)})}{p(\theta_t^{(i)}|\theta_{t-1}^{(i)}, y_t)} && (3.46) \\
&= w_{t-1}^{(i)} p(\theta_t^{(i)}|\theta_{t-1}^{(i)})p(y_t|\theta_t^{(i)}, \theta_{t-1}^{(i)}) \frac{p(y_t|\theta_{t-1}^{(i)})}{p(y_t|\theta_{t-1}^{(i)}, \theta_t^{(i)})p(\theta_t^{(i)}|\theta_{t-1}^{(i)})} && (3.47) \\
&\propto w_{t-1}^{(i)} p(y_t|\theta_{t-1}^{(i)}) && (3.48) \\
&= w_{t-1}^{(i)} \int p(\theta_t|\theta_{t-1}^{(i)})p(y_t|\theta_t)\mathrm{d}\theta_t && (3.49)
\end{aligned}$$

We we substitute the definition of the optimal importance distribution in (3.46); and apply Bayes rule to $p(\theta_t|\theta_{t-1}^{(i)}, y_t)$ in (3.47); cancel terms in (3.48), and substitute the definition of predictive likelihood in equation (3.49) (as introduced in Section 3.2, p. 33). The resulting predictive likelihood is the probability of the observation at time t given the model state at time $t-1$. Computing this involves integrating over all possible parameter values at time t. In models with a discrete and finite state space this is usually possible (the integral becomes a finite sum). We will use the optimal importance function in the particle filters developed in this thesis, which we can do efficiently because we use a collapsed representation of the state space (see Rao Blackwellized particle filtering, Section 3.3.3.4).

In sum, sequential importance sampling in particle filtering proceeds as follows (cf., Algorithm 3 lines 1–8, and our illustration in Figure 3.2). We assume some initial set of particles θ_0. Then at each time t we first update our particle sample by drawing from the recursive importance distribution (which is approximated through as sample as our current set of available particles from time $t-1$). Secondly, we update the particle weights according to (3.44), and finally normalize the weights to sum to one.

Unfortunately, the method as described above tends to be ineffective: even under the optimal importance distribution, the approximation of the target distribution quickly decreases in quality due to the repeated approximations through a limited number of samples from an importance distribution. Practically, a poor approximation manifests in degenerate particle weights: after a few iterations, few or only one particle accumulates the vast majority of particle weight. The set of weighted particles at time t

Algorithm 2 The Particle Filtering Algorithm.

```
 1: for particle $i = \{1, ..., N\}$ do                                    ▷ Initialization
 2:     $\theta_0^{(i)} \sim q_0(\theta)$
 3: for time $t = \{1, ...T\}$ do
 4:     for particle $i = \{1, ..., N\}$ do                 ▷ Sequential Importance Sampling
 5:         sample            $\theta_t^{(i)} \sim q(\theta_{1:t}|\theta_{1:t-1}, \mathbf{y}_{1:t})$
 6:         update samples    $\theta_{1:t}^{(i)} \leftarrow \{\theta_{1:t-1}^{(i)}, \theta_t^{(i)}\}$
 7:         update weights    $\tilde{w}_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(\theta_t^{(i)}|\theta_{t-1}^{(i)})p(y_t|\theta_t^{(i)})}{q(\theta_t^{(i)}|\theta_{1:t-1}^{(i)}, \mathbf{y}_{1:t})}$
 8:     normalize weights $w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_j \tilde{w}_t^{(j)}}$
 9:     if ESS < threshold then                                              ▷ Resampling
10:         draw with replacement $\tilde{\theta}_{1:t} \sim \left\{Multinomial(\mathbf{w}_t)\right\}_{k=1}^{N}$
11:         new sample    $\theta_{1:t} \leftarrow \tilde{\theta}_{1:t}$
12:         re-set weights    $\mathbf{w}_t = \frac{1}{N}$
```

in Figure 3.2 (center) illustrates degenerate particle weights. The posterior distribution is then effectively approximated through a point estimate. We now discuss methods to alleviate this problem.

3.3.3.3 Resampling

Resampling is an effective and widely used method for recovering from a degenerate set of samples (Gordon et al., 1993; Doucet and Johansen, 2008). It can be straightforwardly integrated into sequential importance sampling, leading to the sequential importance resampling (SIR) algorithm. All particle filters derived in this thesis use SIR. Intuitively, whenever the weight variance exceeds a threshold (i.e., few particles have accumulated too much weight), a subset of high-weight particles is probabilistically selected from the full set of particles, and only this set will be propagated further.

Given a set of N particles and their associated weights, a new set of particles is sampled by drawing N times with replacement from a Multinomial distribution parameterized by the particle weights. Any particle which is not sampled in this process 'dies out', and will not be propagated further. After the resampling step, weights are set uniformly to $\frac{1}{N}$, which is valid since the old weights are implicitly represented in the sample. Resampling is displayed in lines 9–12 in Algorithm 2, and illustrated in Figure 3.2.

It is common to define a threshold for acceptable variance among the particle weights, and resample the set of particles whenever this threshold is crossed. A common choice of threshold is the *Effective Sample Size* (ESS) which measures the number of particles which *effectively* contribute to the sample, i.e., have non-negligible weight:

$$\text{ESS} = \frac{1}{\sum_i (w_t^{(i)})^2}, \tag{3.50}$$

where ESS decreases with increasing variance among the particle weights. Whenever the ESS falls below a threshold, a resampling step is executed.

Resampling allows to replace low-weight particles with high-weight particles by repositioning the sample in high probability space under the posterior distribution. However, it introduces additional noise to the sampling process. After all, samples from the posterior distribution might be pruned which seem poor at time t but may become more fitting in the future after more data was observed. Furthermore, it leads to copies of identical particles being propagated. While focusing the sampler on high-probability areas under the target distribution, it does so at the cost of diversity in the sample. Resampling can consequently result in an *impoverished* set of samples.

Rejuvenation Sample impoverishment can be minimized by keeping resampling steps to a minimum, e.g., by choosing an appropriate threshold for the effective sample size. In addition, impoverished samples can be improved by ‘disturbing’ each resampled particle slightly, enhancing the diversity in the set of resampled particles. This approach is called *rejuvenation*, and is part of widely used variants of particle filters such as the resample-move algorithm (Gilks and Berzuini, 2001). Rejuvenation is illustrated in the bottom part of Figure 3.2. Immediately after resampling, a limited number of MCMC steps are executed individually within each resampled particle. The MCMC sampler (e.g., a Gibbs sampler) is constructed such that its stationary distribution corresponds to the target distribution of the particle filter. Consequently, after rejuvenation the particles are still a valid sample from the target distribution. We use rejuvenation as described above in the particle filters developed in this thesis.

3.3.3.4 Rao-Blackwellized Particle Filtering

Some models allow to analytically integrate over subsets of their parameters. We discussed this marginalization in the context of Dirichlet-Multinomial models (Section 3.2.2.2), and the collapsed Gibbs sampler (Section 3.3.2.3). The same idea can be used with particle filters, which results in Rao-Blackwellized particle filters. Rao-Blackwellized particle filters sequentially estimate only the remaining model parameters, which cannot be marginalized (Liu and Chen, 1998; Doucet et al., 2000a).

The Rao-Blackwellized particle filter is generally advantageous to use because it operates on a reduced state space which has been shown to lead to improved efficiency and robustness (Liu and Chen, 1998; Doucet et al., 2000b). Rao-Blackwellized particle filters have been employed for incremental clustering problems which are similar to those discussed in this thesis (Sanborn et al., 2006; Canini et al., 2009).

We use Rao-Blackwellized particle filtering throughout this thesis. Intuitively, our particle filters will incrementally assign discrete latent labels (e.g., categories) to observations over time. The continuous parameters underlying the Multinomial distributions in our models are collapsed, i.e., not estimated explicitly, but implicitly represented through their sufficient statistics.

3.4 Summary

In this chapter we reviewed the mathematical background underlying the models developed in this thesis. We began by motivating Bayesian modeling as a framework for computational investigations of cognitive phenomena which formulates inductive inference under uncertainty in a mathematically principled way. We also motivated sampling-based approximate Bayesian inference as a flexible and general method to explore the processes and limitations underlying human cognition. The second part of the chapter formally introduced Bayesian statistical modeling, discussed prior distributions and likelihood functions relevant to the models of this thesis, and demonstrated their characteristics and limitations. The final part introduced the Monte Carlo method in the context of approximating the posterior distribution of hierarchical Bayesian models. Two concrete instantiations were introduced: a Gibbs sampler, which uses the Markov chain Monte Carlo technique and operates in a batch fashion, and a particle filter, which approximates the posterior distribution sequentially.

In the following chapters will introduce cognitively motivated models of category learning and meaning development which make use of the theoretical framework outlined in this chapter.

Chapter 4

Incremental Bayesian Category Learning

The task of *categorization*, in which people cluster stimuli into categories and then use those categories to make inferences about novel stimuli, has long been a core problem within cognitive science. Understanding the mechanisms involved in categorization, particularly in category acquisition, is essential, as the ability to generalize from experience underlies a variety of common mental tasks, including perception, learning, and the use of language. As a result, category learning has been one of the most extensively studied aspects in human cognition, both from an empirical and modeling perspective. In a typical experiment, participants are taught the category membership of a set of training stimuli and then asked to generalize to a set of test stimuli. Computational models are then evaluated on their ability to predict the resulting patterns of generalization (Anderson, 1991).

Categorization is a classic example of inductive inference, i.e., extending knowledge from known to novel instances. When learning about a new category of objects, humans need to infer the structure of the category from examples of its members. The knowledge acquired through this process can ultimately be used to make decisions about how to categorize new stimuli. Categorization presents a difficult inference problem: the learner is faced with limited data (e.g., a few concept observations), and has to evaluate several categorization hypotheses given this data without knowing exactly which category structure is correct. Furthermore, inference proceeds *incrementally*, learners encounter data and update their beliefs over time, making new generalizations when new information becomes available (Bornstein and Mash, 2010; Diaz and Ross, 2006). To complicate matters, categorization is an example of a *joint* inference problem. For instance, experimental evidence suggests that the development of categories and their characteristic features emerge simultaneously in one process (Goldstone et al., 2001; Schyns and Rodet, 1997). It is also well-known that children's word learning improves when they form some abstract knowledge about what kinds of semantic properties are relevant to what kinds of categories (Jones et al., 1991; Colunga and Smith, 2005; Colunga and Sims, 2011). This abstract knowledge is argued to emerge by generalizing over the learned words. So, words that have been learned contribute to generalized abstract knowledge about word meanings and semantic categories, which then guide subsequent word learning.

In this chapter, we present a computational model which tackles the problem of learning categories and their characteristic features from natural language text. Our model is presented with concepts such as {*parrot, seagull, chocolate, sausage*} and their local context, and groups them into categories (BIRD and FOOD in this example) based on their contextual similarity. Although concepts like *parrot* and *seagull* might rarely co-occur together explicitly, they do occur in similar contexts (e.g., {`croak,lay-eggs`}[1]). Analogously, the concepts *chocolate* and *sausage* might rarely be observed together in text, however, they share contexts such as {`eat,breakfast,healthy`}. We thus approximate category-specific features with natural language context, and show that our model learns meaningful categories as well as descriptive features for them.[2] More technically, our model of category acquisition is based on the key idea that learners can adaptively form category representations that capture the structure expressed in the observed data. We model category induction as two interrelated sub-problems: (a) the acquisition of features that discriminate among categories, and (b) the grouping of concepts into categories based on those features. Our model learns *incrementally* as data is presented and updates its internal knowledge state locally without systematically revising everything known about the situation at hand.

We formulate our categorization model in a probabilistic Bayesian setting. Probabilistic approaches provide a computational framework for modeling inductive problems,

[1] Throughout this thesis we will use small caps to denote CATEGORIES, italics to denote their *members*, and typewriter fonts for their `features`.

[2] We use the terms concepts and categories to refer to *basic level* and SUPERORDINATE categories, respectively. Our model in turn infers superordinate categories based on the features of their basic level category members.

by identifying ideal or optimal solutions to them and then using algorithms for approximating these solutions (cf., Section 3.1 for an extended discussion of Bayesian cognitive modeling). Several probabilistic category learning models have been proposed in the literature (Anderson, 1991; Ashby and Alfonso-Reese, 1995; Griffiths et al., 2008; Sanborn et al., 2010; Canini, 2011), essentially viewing category learning as a problem of density estimation: determining the probability distributions associated with different category labels. Our model learns categories using a particle filter (Doucet et al., 2001), a sequential Monte Carlo (SMC) inference mechanism which allows to update a probability distribution over time, while sequentially integrating newly observed data. Section 3.3.3 contains a technical introduction to particle filters. Monte Carlo algorithms offer a plausible proxy for modeling human learning and have been previously used (Börschinger and Johnson, 2011, 2012; Levy et al., 2009; Sanborn et al., 2010; Griffiths et al., 2008) to explain how humans might be performing probabilistic inference, essentially reducing probabilistic computations to generating samples from a probability distribution.

Historically, the stimuli involved in categorization studies (either laboratory experiments or computational simulations) tend to have a small number of manually specified features, and are either concrete objects (such as physical objects, Bornstein and Mash 2010) or highly abstract ones (such as binary strings, colored shapes, Medin and Schaffer 1978; Kruschke 1993; Lee and Navarro 2002). Most existing models focus on adult categorization, in which it is assumed that learners have developed categorization mechanisms and a large number of categories have already been learnt. Those models are typically evaluated against behavioral data elicited in laboratory experiments from adult participants who are assumed to have acquired and are able to make use of rich prior world knowledge. A notable exception is Anderson's (1991) rational model of categorization (see also Griffiths et al. 2007a) where the learner starts without any predefined categories and stimuli are clustered into groups as they are encountered. Our model is based on the same assumption (i.e., it learns categories directly from data), but instead uses natural language stimuli (i.e., words).

The idea of modeling categories using words as a stand-in for their referents has been previously used to explore categorization-related phenomena such as semantic priming (Cree et al., 1999) and typicality rating (Voorspoels et al., 2008), to evaluate prototype and exemplar models (Storms et al., 2000), and to simulate early language category acquisition (Fountain and Lapata, 2011). The idea of using naturalistic corpora as a proxy for people's representation of semantic concepts has received little attention. Instead, featural representations, called feature norms, have played a central role in psychological theories of semantic cognition and knowledge organization and many studies have been conducted to elicit detailed knowledge of features (Smith et al., 1974; McRae et al., 2005; Vinson and Vigliocco, 2008; Rogers and McClelland, 2004). In a typical procedure, participants are presented with a word and asked to generate the most relevant features or attributes for its referent concept (e.g., McRae et al. 2005). Our approach replaces feature norms with representations derived from words' contexts in corpora. We assume that words whose referents exhibit differing features are likely to occur in correspondingly different contexts and that these differences in usage can provide a substitute for featural representations.

While this is an impoverished view of how categories are acquired – it is clear that they are learnt through exposure to the linguistic environment *and* the physical world – perceptual information relevant for extracting semantic categories is to a large extent redundantly encoded in linguistic experience (Riordan and Jones, 2011). Besides, there are known difficulties with feature norms such as the small number of words for which these can be obtained, the quality of the attributes, and variability in the way people generate them (see Zeigenfuse and Lee 2010 for details). Focusing on natural language categorization allows us to build models with theoretically unlimited scope. Moreover, the corpus-based approach is attractive for modeling the *development* of linguistic categories. If simple distributional information really does form the basis of a word's cognitive representation (Harris, 1954; Redington and Chater, 1997; Braine, 1987), this implies that learners are sensitive to the structure of the linguistic environment during language development. As experience with a word accumulates, more information about its contexts of use is encoded, with a corresponding increase in the ability of the language learner to use the word appropriately and make inferences about novel words of the same category.

In the remainder of this chapter, we review previous research on categorization placing emphasis on natural language categories and Bayesian models. Next, we present our categorization model and its incremental learning mechanism, and describe several experiments assessing its performance when applied to a large corpus as well as to a smaller corpus of child-directed speech. Experimental results show that our incremental learner obtains meaningful categories which yield a closer fit to behavioral data compared to related models whilst at the same time acquiring features which characterize the learnt categories. In all cases, we evaluate the induced categories by comparing model output against a gold standard set of categories and concepts created by humans.

4.1 Category Learning from Natural Language

Numerous theories as to how humans categorize objects have been proposed and extensively tested, and here we highlight those relevant to our modeling approach. *Prototype* theory (Rosch, 1973) represents categories through an idealized prototypical member possessing the features which are critical to the category. Membership in the category is determined by comparing the observed features of a possible member against those of the prototype. For example, the characteristic features of FRUIT might include `contains seeds`, `grows above ground`, and `is edible`.

Prototype theory has been challenged by the *exemplar* approach (Medin and Schaffer, 1978). In this view, categories are defined not by a single representation but rather by a list of previously encountered members. An exemplar model simply stores those instances of fruit to which it has been exposed (e.g., *apples*, *oranges*, *pears*). A new object is grouped into the category if it is sufficiently similar to one or more of the FRUIT instances stored in memory. Practically, exemplar models and prototype models can account for the same range of phenomena. Our Bayesian model of categorization resembles an exemplar model: information from all encountered concept observations is stored and contributes to the representation of their particular category.

The *knowledge* approach to categories takes a somewhat different standpoint asserting that categories are formed on the basis of people's general knowledge about the world. This view is perhaps best illustrated by what Barsalou (1985) calls goal-derived categories, i.e., categories that are designed based on how their members fill some externally-determined role. For example, the category of BREAKFAST FOODS, consisting of concepts like *bacon*, *eggs*, or *grits* is quite clearly a category people can and do form, and about which they can make meaningful judgments, yet there is very little similarity between members, making it difficult to account for using an exemplar model or a prototype model. Our own model learns from large corpora which can be viewed as a rich source of world knowledge. It makes use of the knowledge encoded in a a word's context to form abstractions that are qualitatively different from those that can be encapsulated by either exemplars or prototypes. We show in our experiments that the kinds of categories and features our model induces are representative of background knowledge.

Models and Modalities of Language Acquisition In this work we formulate a categorization model which learns from exposure to the distributional properties of the linguistic environment. However, it is clear that when children learn language, they are not only exposed to linguistic input but also to various types of perceptual input, including visual context, prosody, gaze and body movement. Additionally, learning is cross-situational – children learn words or concepts through repeated co-occurrence of clues from different modalities in the environment (such as objects and their linguistic labels) – which implies that learners combine information from both linguistic and nonlinguistic context. Here, we briefly overview the ways in which various modalities have been incorporated in computational models of language acquisition, and position our own model in the context of this work. A more thorough discussion of this line of prior work is presented in Chapter 2.

A variety of models on cross-modal word learning have been proposed. Word learning is the process of creating a “mental lexicon” from linguistic input, identifying words and their referents, and as such is a form of categorization. These models range from combining raw speech with visual input (Roy and Pentland, 2002), or concrete objects with words (Xu and Tenenbaum, 2007), to eliciting cross-situational co-occurrence patterns of linguistic input and objects in speakers’ attention (Frank et al., 2009).

Acquisition of visual categories is an important and notoriously hard problem in the area of computer vision, where large-scale systems require thousands of training examples with sophisticated features in order to be able to recognize classes of objects in images. This stands in sharp contrast to humans who quickly and robustly recognize objects regardless of scale or perspective. Fei-Fei et al. (2003) propose a Bayesian model for category learning from purely visual image data incorporating prior knowledge in the model and show that information based on previously acquired categories boosts learning of new categories.

Another line of work investigates the joint process of word learning and object categorization showing that linguistic cues facilitate object recognition and vice versa (see also Lupyan et al. 2007). Yu (2005) develops a joint model of lexical acquisition and object categorization based on experimental evidence indicating that the two problems are interrelated. The model learns from linguistic and visual data (simplified as color, shape and texture features). Specifically, subjects were asked to narrate a picture book wearing a head-mounted camera to capture a first-person point of view while their acoustic signals were being recorded (using a headset microphone). Similarly, Yu and Ballard (2004) simulate joint word and object learning in adults based on descriptions of nine objects paired with images from a head-mounted camera.

The models introduced above require complex and controlled multimodal input data, which inherently limits their scope. While their aim is to support fundamental characteristics of language acquisition it is unclear whether the models generalize to other tasks or types of data. In this work we adopt a complementary approach. While we consider a qualitatively coarser approximation of the learning environment, in the form of linguistic corpora, this has the advantage of being able to test our models on a larger scale. Below, we discuss our approach in more detail contrasting it to related work focusing exclusively on categorization.

Natural Language Categorization Most experimental work on category modeling and acquisition has revolved around laboratory experiments involving either real-world objects (e.g., children's toys; Starkey 1981), perceptual abstractions (e.g., photographs of animals; Quinn and Eimas 1996), or abstract, artificial stimuli (e.g., dot patterns or geometric shapes; Posner and Keele 1968 and Bomba and Siqueland 1983, respectively). In most cases researchers using abstract or artificial stimuli to explore human categorization would not assert that participants possess a distinct mechanism for distinguishing between categories of (for example) binary strings, but rather that the task invokes a single, global mechanism for learning and applying categories. Our own approach is no different, in that we treat word meaning as a proxy for conceptual structure (Murphy, 2002) and do not suggest that (semantic) categories of words differ significantly from the categories involving their real-world referents. We refer to this task, of organizing words into categories based on their semantics, as *natural language categorization*. While the idea of modeling categories using words as a stand-in for their referents is of course not a new one, explicitly viewing categorization as the task of organizing words into categories based on meaning allows us to make use of powerful ideas from artificial intelligence and computational linguistics. Previous work that could be described as natural language categorization has a recurring theme: the use of feature norms to construct semantic representations for word meaning. Feature norms are traditionally collected through norming studies, in which participants are presented with a word and asked to generate a number of relevant features for its referent concept (The most notable of these is probably the multi-year project of McRae et al. (2005), which collected and analyzed features for a set of 541 common English nouns). The results of such studies can be interesting in their own right, as the frequency and distribution of generated features can provide considerable insight into the nature of participants' categories — but they can also provide material for evaluating prototype and exemplar models.

Existing research into natural language categorization has used such featural representations to explore a wide range of categorization-related phenomena. Heit and Barsalou (1996) demonstrated their instantiation principle within the context of natural language concepts, Storms et al. (2000) contrasted exemplar and prototype models using a task-based evaluation, Cree et al. (1999) used feature-based representations to model semantic priming, and Voorspoels et al. (2008) model typicality ratings for natural language concepts. In all of these models words are used as a proxy for real-world stimuli, and feature norms as a proxy for people's perceptual experiences of those stimuli. Our approach is to replace feature norms with representations derived from words' context in corpora, i.e., to use distributional semantics to approximate people's perceptual representations of real-world stimuli. While this approach represents only a partial view of how people acquire and use categories, experimental comparisons of feature-based and corpus-based categorization models indicate that the latter represent a viable alternative to the feature norms typically used (Fountain and Lapata, 2010).

Our work is closest to Fountain and Lapata (2011) who also develop a corpus-based model of natural language categories drawing inspiration from semantic networks (Collins and Loftus, 1975). In this framework, each node is a word, representing a concept (like BIRD). With each node is stored a set of properties (like `can fly` or `has wings`) as well as links to other nodes (like CHICKEN). A node is directly linked to those nodes of which it is either a subclass or superclass (i.e., BIRD would be connected to both CHICKEN and ANIMAL). High-level nodes representing large categories are connected (directly or indirectly) to many instances of those categories, whereas nodes representing specific instances are at a lower level, connected only to their superclasses. A word's meaning is expressed by the number and type of connections it has to other words. Semantic networks constitute a somewhat idealized representation that abstracts away from real word usage. The model on its own does not specify how the representations are learned and the latter are traditionally hand-coded by modelers who have to *a priori* decide which relationships are most relevant in representing meaning.

The model presented in Fountain and Lapata (2011) is *distributional*, i.e., it represents the meaning of words by their patterns of co-occurrence with other words. They also organize concepts in a semantic network that is not, however, structured hierarchically. They consider a simpler formulation of semantic networks in which a network is composed of a graph with edges between word nodes. Such a graph is *unipartite*: there is only one type of node, and those nodes can be interconnected freely. Edges between nodes do not represent subsumption but similarity or relatedness and can be easily quantified in a distributional framework (words that are similar in meaning will tend to behave similarly in terms of their distributions across different contexts). Their model is an incremental version of Chinese Whispers (Biemann, 2006), a randomized graph-clustering algorithm. The latter takes as input a graph which is constructed from corpus-based co-occurrence statistics and produces a hard clustering over the nodes in the graph. Their model treats the tasks of inferring a semantic representation for concepts and their class membership as two separate processes. This allows to experiment with different ways of initializing the co-occurrence matrix (e.g., from bags of words or a dependency parsed corpus), however at the expense of cognitive plausibility. It is unlikely that humans have two entirely separate mechanisms for learning the meaning of words and their categories. We formulate a more expressive model which captures word categories and their predictive features in one, unified process.

Bayesian Models Incremental Bayesian category learning was pioneered by Anderson (1991) who developed a non-parametric model able to induce categories from abstract stimuli represented by binary features. According to this model, category learning amounts to Bayesian density estimation, where the number of clusters to be used in representing a set of objects is selected automatically. Sanborn et al. (2006) and Sanborn et al. (2010) present a fully Bayesian adaptation of Anderson's original model, which yields a better fit with behavioral data. Specifically, borrowing ideas from nonparametric Bayesian statistics, they propose two algorithms for approximate inference in this model: Gibbs sampling (a "batch" procedure where density estimation assumes that all data are available at the time of inference) and particle filtering (where density estimation proceeds incrementally over time, as stimuli become available). A separate line of work examines the processes of generalizing and generating new categories and concepts (Jern and Kemp, 2013; Kemp et al., 2012) which are again modeled as samples from probability distributions.

In this work, we also present a probabilistic Bayesian model of categorization which is conceptually similar to Sanborn et al. (2010). However, our model was developed with (early) language acquisition in mind. They focus on adult categorization and use rather simplistic categories representing toy-domains. It is therefore not clear whether their approach generalizes to arbitrary stimuli and data sizes. Moreover, they are primarily interested in how to approximate the intractable ideal solution to the partitioning problem. Our work differs in two respects: firstly, we are interested in large-scale categorization. We investigate the question whether it is possible to learn categories from a large number of observations of concepts covering a wide variety of categories, thus approaching the scale of the problem that a child is faced with. Secondly, we are interested in learning the representations for real-world, semantic categories of concrete, observable objects (for example, that a *dog* is an ANIMAL or that a *chair* is FURNITURE).

Latent Dirichlet Allocation (LDA; Blei et al. (2003)) is a popular Bayesian model for discovering latent topics in text. LDA assumes that a document is generated from an individual mixture over topics, and each topic is characterized by a distribution over words. LDA learns topics from longer documents whereas we argue that a limited *local* context is appropriate for category induction since a target concept's features are best represented through its immediately surrounding words. Fountain and Lapata (2011) further show that LDA cannot be applied effectively to shorter contexts appropriate for category acquisition. From a cognitive point of view, focusing on local contexts of target concepts approximates limitations of attention and memory faced by young learners. Finally, it is unclear how to naturally define longer contexts when the input given to the model consists of streams of child-directed speech. Our model infers a grouping of words into semantic categories based on the assumption that local linguistic context can provide important cues for word meaning and by extension category membership. In this sense, it is loosely related to Bayesian models of word sense induction (Brody and Lapata, 2009; Yao and Durme, 2011) which also make use of short local contexts. However, the above models focus on performance optimization and learn in an ideal batch mode, while incorporating various kinds of additional features such as part of speech tags or syntactic dependencies. In contrast, we develop a cognitively plausible (early) language learning model and show that categories can be acquired purely from linguistic context, as well as in an incremental fashion.

From a modeling perspective, we learn categories using a particle filtering algorithm (Doucet et al., 2001). As explained in Section 3.3.3, Particle filters are a family of sequential Monte Carlo algorithms which update the state space of a probabilistic model with newly encountered information. Particle filters have been previously used to explain behavioral patterns in several tasks such as associative learning (Daw and Courville, 2007), change-point detection (Brown and Steyvers, 2009), word segmentation (Börschinger and Johnson, 2011), and sentence processing (Levy et al., 2009). As mentioned earlier, Sanborn et al. (2006) also use particle filters for small-scale categorization experiments with artificial stimuli. To the best of our knowledge, we present the first particle filtering algorithm for large-scale category acquisition from natural language text.

4.2 Bayesian Natural Language Categorization

We begin by formalizing the general problem of Bayesian categorization and then derive our model as an instance of this formulation. In this framework, the learner is faced with a partitioning problem, i.e., to group observed concepts into categories based on their features. We use the term *stimuli* to denote linguistic observations of concepts and their features. A common assumption is that concepts with sufficiently similar features will be assigned to the same category. During this learning process, categories are not directly observed but are instead inferred from their observable features. Once categories are established, the learnt category-specific features can be used to predict the category of new concepts.

More formally, given a stimulus d, a Bayesian model of categorization predicts a latent category z_d based on the observable features x_d of the stimulus, as well as the information observed from previously encountered stimuli $\mathbf{x}_{d-1}$, and the latent category assignment $\mathbf{z}_{d-1}$. Based on this information, we compute for stimulus d the probability of being assigned category j:

$$P(z_d = j | x_d, \mathbf{z}_{d-1}, \mathbf{x}_{d-1}) = \frac{P(z_d = j | \mathbf{z}_{d-1}) \times P(x_d | z_d = j, \mathbf{x}_{d-1}, \mathbf{z}_{d-1})}{\sum_{j'=1}^{J} P(z_d = j' | \mathbf{z}_{d-1}) \times P(x_d | z_d = j', \mathbf{x}_{d-1}, \mathbf{z}_{d-1})}. \quad (4.1)$$

The Bayesian formulation of this problem computes the posterior probability of the category assignment $P(z_d = j)$ based on two factors. The first term of the numerator in equation (4.1) is the prior probability of selecting category j based on the category assignments of the previously assigned concept observations. A common choice for this prior is a 'rich-get-richer' scheme: categories which have been chosen frequently in the past, are more likely to be selected again. The second term of the numerator in equation (4.1) is the likelihood term, which considers x_d, the observed features of stimulus d, and computes the probability that they were generated from category j. By assigning each stimulus to exactly one category, the learning process discovers a partition of stimuli into categories consistent with the observable data. In order to find the optimal partitioning, it would be necessary to iterate over all possible partitionings of the data, which is intractable for any data set of non-trivial size. Several approximation algorithms for this problem have been proposed, one of which, namely particle filtering, we will describe later in this section.

The model presented above is very general and as such can be applied to many different types of stimuli and features. For example, Sanborn et al. (2010) (following Medin and Schaffer 1978) use a small number of artificial stimuli, each with four binary features (e.g., 1111, 0101, 1010). In another experiment, they use 12 stimuli with continuous features, varying in brightness and saturation. Other work focusing on natural language categorization has assumed that abstract cognitive representations of concepts can be represented as sets of features obtained from norming studies. Table 4.1 (top) provides examples of concepts and their elicited features.

In our work we learn the semantic representations of concepts from large-scale linguistic corpora without relying on explicit human judgment. In this framework, information about the meaning of words can be derived by analyzing the co-occurrences between words and the contexts in which they occur. Many cognitive models of word meaning (Landauer and Dumais, 1997; Griffiths et al., 2007b; Lund and Burgess, 1996) subscribe to this distributional hypothesis which states that a word's meaning is predictable from its context (Harris, 1954). By extension, we further assume that a word's context is predictive of its category and that category features can be derived from the linguistic context. Our model (incrementally) *learns* semantic categories based on the linguistic features of their context, and can be tested on a large scale. Table 4.1 (bottom) shows examples of the linguistic features we consider for different concepts.

		strawberry	grape	apple	snail	dog	cat
Feature Norms	has_a_taste	✓	✓	✓			
	contains_seeds	✓	✓	✓			
	is_edible	✓	✓	✓			
	can_be_a_pet				✓	✓	✓
	is_alive	✓	✓	✓	✓	✓	✓
	eats				✓	✓	✓

		strawberry	grape	apple	snail	dog	cat
Context Features	*ripe*	✓	✓	✓			
	hungry	✓	✓	✓	✓	✓	✓
	lemon	✓	✓	✓			
	owner				✓	✓	✓
	bark					✓	
	shepherd					✓	✓

Table 4.1: Concepts and their features for the categories FRUIT and ANIMAL. Features are shown as feature norms (top) and as context words (bottom).

4.3 A Bayesian Model of Large-scale Incremental Category Learning

In this section we present our Bayesian model for large-scale semantic category acquisition from natural language text (BayesCat for short). For now we focus on the *computational* level (Marr, 1982) of the problem definition of categorization, and present a model with which we can (in principle) learn semantic categories. In the following section we turn to the *algorithmic* dimension of the problem, and introduce two learning algorithms for our model: a batch algorithm, which learns by repeated iterations over the entire training data set (Section 4.3.1); and a more cognitively plausible incremental inference algorithm which accumulates information in real time, as stimuli are observed (Section 4.3.2).

Intuition The input BayesCat receives is natural language text, and its final output is a set of categories (aka clusters) as discovered from the input stimuli. We use the linguistic context of observed concepts as a proxy for their characteristic features, and assume that concepts with sufficiently similar features are assigned to the same category. The model is exposed to linguistic stimuli, each consisting of a target concept t and a set of context words c from a symmetric window of length n:

$$[c_{-n} \dots c_{-1} \;\; t \;\; c_1 \dots c_n]. \tag{4.2}$$

Each induced category will be characterized by a set of concepts which are members of the category, as well as a set of category-specific features. We assume a global distribution over categories θ, from which all stimuli are generated. Each category k has two associated multinomial distributions over words: (1) a distribution over concepts (i.e., target words) ϕ_k and (2) an independently parametrized distribution over context words ψ_k. The separation of concepts from context words allows us to learn features together with category members. We furthermore argue that, while members of the same category tend to appear in the same contexts, they do not necessarily co-occur. For example, the concepts *parrot* and *seagull* are both members of the category BIRD, but are rarely mentioned together, however, they frequently occur with the same features, e.g., they both `fly`, `croak`, `lay eggs`, and so on.

Model Description A graphical overview of the BayesCat model in form of a plate diagram is shown in Figure 4.1b. Figure 4.1a displays the generative process of the BayesCat model which proceeds as follows.[3] First, we draw parameters θ for a global distribution over categories from a Dirichlet distribution with parameter α. Then, for each category k, we draw (1) parameters ϕ_k for a category-specific concept distribution (from a Dirichlet distribution with parameter β), as well as (2) parameters ψ_k for a category-specific context word (or feature) distribution (from a separate Dirichlet distribution parametrized by γ). Using these global parameters, we can generate stimuli d. First, draw a category $z^d \sim Mult(\theta)$. Then, draw a target word from the category-specific concept distribution $w_t^d \sim Mult(\phi_{z^d})$; and finally, independently for each context position i, we draw a context word from the category-specific feature distribution $w_c^{d,i} \sim Mult(\psi_{z^d})$.

The full joint distribution over data and model parameters as defined by our model (see the independence assumptions in the plate diagram in Figure 4.1b) can be factorized

[3] We refer to the Dirichlet distribution as *Dir* and to the Multinomial distribution as *Mult*.

(a) Generative story of BayesCat.

```
Draw distribution over categories $\theta \sim Dir(\alpha)$
for category $k$ do
    Draw target word distribution $\phi_k \sim Dir(\beta)$
    Draw context word distribution $\psi_k \sim Dir(\gamma)$
for stimulus $d$ do
    Draw category $z^d \sim Mult(\theta)$
    Draw target word $w_t^d \sim Mult(\phi_{z^d})$
    for context position $i = \{1...I\}$ do
        Draw context word $w_c^{d,i} \sim Mult(\psi_{z^d})$
```

(b) Plate diagram of BayesCat.

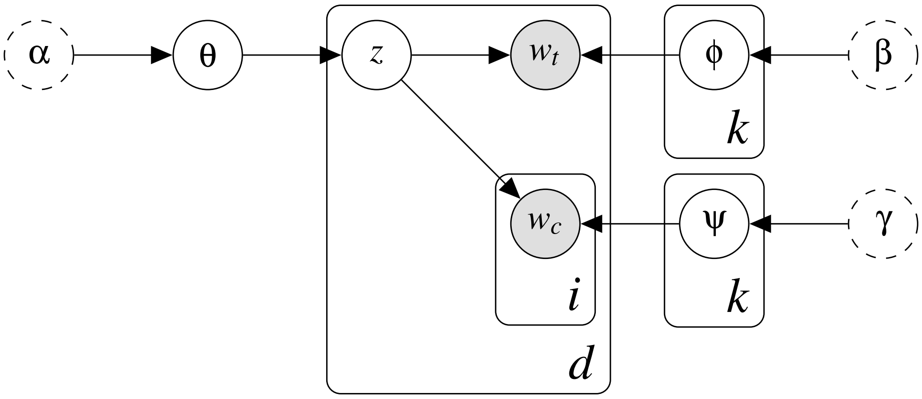


Figure 4.1: Top (a): The generative story of the BayesCat model. Observations (w_t and w_c) and latent labels (z) are drawn from Multinomial distributions ($Mult$). Parameters for the multinomial distributions are drawn from Dirichlet distributions (Dir). Bottom (b): The plate diagram representation of the BayesCat model. Observed variables (target concepts and context words) are shown as shaded nodes, white solid nodes represent the latent variables to be estimated, and fixed hyper-parameters are shown as white dashed nodes. Plates indicate repetition of the variables they contain with the subscript indicating the number of repetitions (e.g., the model contains an individual distribution over concepts ϕ for each category k).

as:

$$
\begin{aligned}
&P(\mathbf{y},\mathbf{z},\theta,\phi,\psi;\alpha,\beta,\gamma) = \\
&\qquad P(\theta|\alpha) \times \prod_{k=1}^{K} P(\phi_k|\beta)P(\psi_k|\gamma) \times \prod_{d=1}^{D} P(z^d|\theta)P(w_t^d|\phi_{z^d}) \prod_{i=1}^{I} P(w_c^{d,i}|\psi_{z^d}),
\end{aligned}
\tag{4.3}
$$

where $\mathbf{y}$ refers to all observed data, $\mathbf{z}$ refers to the hidden category labels, and k,d and i are indices ranging over categories, stimuli, and context positions, respectively. The parametrization of our model allows us to further simplify the joint distribution. Due to the conjugacy of the Dirichlet and Multinomial distribution, we can analytically integrate over all possible values of the model's parameter distributions θ,ϕ and ψ (see Section 3.2.2 for the technical details). Dirichlet-Multinomial distributions encode a "rich-get-richer" scheme: if a category has been frequently assigned to previously encountered stimuli, it is more likely that it will be observed again. Intuitively, this triggers learning of multinomial parameters which distribute most of their mass over few words, i.e., inferring a targeted vocabulary for each individual category. The simplified posterior distribution is:

$$
\begin{aligned}
&P(\mathbf{y},\mathbf{z},\theta,\phi,\psi;\alpha,\beta,\gamma) \propto \\
&\qquad \frac{\prod_k \Gamma(n_k+\alpha_k)}{\Gamma(\sum_k n_k+\alpha_k)} \times \prod_{k=1}^{K} \frac{\prod_r \Gamma(n_r^k+\beta_r)}{\Gamma(\sum_r n_r^k+\beta_r)} \times \prod_{k=1}^{K} \frac{\prod_s \Gamma(n_s^k+\gamma_s)}{\Gamma(\sum_s n_s^k+\gamma_s)},
\end{aligned}
\tag{4.4}
$$

where r ranges over target concepts, s ranges over context words (or features), and $\Gamma(\cdot)$ is the Gamma function. Note that the model parameter distributions do not appear on the right-hand side of equation (4.4). Instead, the model is represented purely through occurrence counts of categories n_k as well as co-occurrence counts of categories with observed concepts and features, n_r^k and n_s^k, respectively. For the interested reader, we derive this result, in Appendix A.

Having motivated and derived a cognitive model for inferring semantic categories from natural text, we now turn to the problem of how these categories are actually learnt (Marr's (1982) *algorithmic* level of analysis) and introducing two learning mechanism. Equation (4.3) defines a probability distribution over all possible partitionings of the concept observations into categories. Exact computation of this density is both computationally intractable an cognitively implausible. It is unrealistic to assume that human learners perform optimal inference (Sanborn et al., 2010). Memory limitations prevent them from enumerating extraordinarily high numbers of hypotheses. Additionally, they make mistakes during learning, and often revisit past decisions in the light of new information. Intuitively, the BayesCat model must *approximate* the target posterior density over all possible partitionings of the concept observations.

We now derive two sampling-based approximate learning algorithms for the BayesCat model, a batch learner (Gibbs sampler; Section 4.3.1), and a cognitively more plausible incremental learner (particle filter; Section 4.3.2).

4.3.1 Batch Learning

We derive a Gibbs sampler for learning the parameters of the BayesCat model in a batch fashion. Gibbs sampling (Geman and Geman, 1984) is a Markov chain Monte Carlo technique for approximating complex joint probability distributions (see Section 3.3.2 for a technical introduction). It operates in batch-mode by repeatedly iterating through all data points (linguistic stimuli in our case) and assigning the currently sampled document d a category z^d conditioned on the current labelings of all other documents z^{-d}:

$$z^d \sim P(z^d|z^{-d}, W^{-d}; \alpha, \beta, \gamma), \tag{4.5}$$

using equation (4.4) but ignoring information from the currently sampled document in all co-occurrence counts.

The Gibbs sampler can be seen as an ideal learner, which can access and revise any relevant information at any time during learning. From a cognitive perspective, this setting is implausible. Humans do not learn in a "batch" fashion, repeatedly and systematically revisiting all information available. Instead, they update their beliefs or knowledge state over time, drawing inferences every time new information arrives. Category learning is no exception and indeed experimental evidence suggests that both children and adults learn categories incrementally (Bornstein and Mash, 2010; Diaz and Ross, 2006).

4.3.2 Incremental Learning

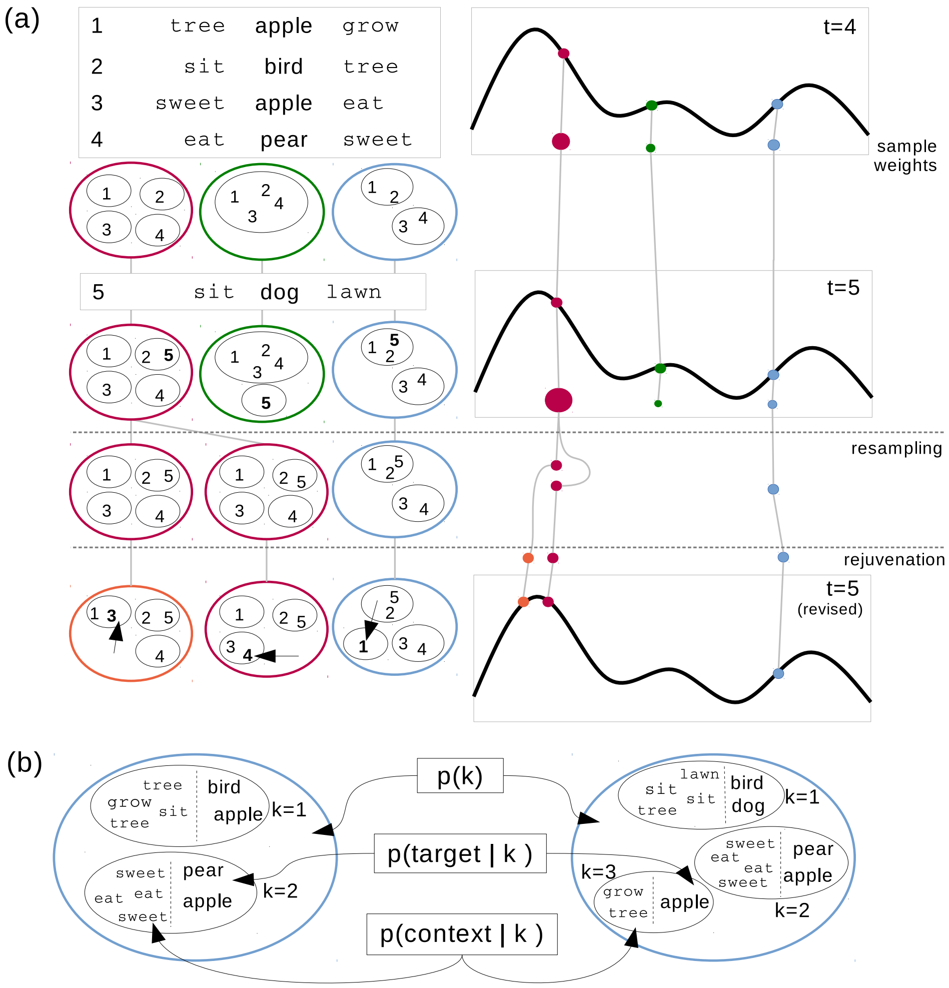
Particle filters are a class of incremental, or sequential, Monte Carlo methods which can be used to model aspects of the human learning process more naturally. The particle filter approximates the target posterior density over all possible partitionings of the concept observations through a set of samples in an *incremental* fashion. Each sample will correspond to one possible categorization of the observed concepts, and each sample will be individually and incrementally updated with information from newly observed stimuli. As is the case in human categorization, the computation time of the updates must stay fixed irrespectively of the number of previously observed concepts. We achieve this by committing to past categorization decisions made by the learning algorithm, and thus integrate a new concept observations *given* the category assignments of all previously encountered concepts (however, we will relax the strict incrementality assumption in the following section).

In the following section we formally describe our learning algorithm, and illustrate it schematically using the example in Figure 4.2a. The full incremental algorithm is displayed in Algorithm 3. A technical introduction to the principles underlying particle filtering can be found in Section 3.3.3 of this thesis.

4.3.2.1 A Particle Filter for the BayesCat Model

Incremental inference algorithms are designed to update estimates of the target distribution with new data becoming available over time. Incremental Monte Carlo algorithms in particular propagate a set of N hypotheses, or samples (called particles) through time and update them with new information. We introduce time into our learning process by treating the observation of each stimulus as one time point. In the example in Figure 4.2a, we show the learning update at time point 4, i.e., after the model has observed stimuli 1–4. The algorithm performs one iteration over the complete set of input stimuli. Our algorithm is based on sequential importance sampling (SIS; Gordon et al. 1993), where the true target distribution is approximated through a simpler *importance* distribution, and the discrepancy between the distributions is counterbalanced through a weight (called importance weight) which is assigned to each sample. A technical introduction to particle filtering and sequential importance sampling can be found in Section 3.3.3.

During learning, we incrementally approximate the target density, i.e., the probability distribution over all possible categorizations of all concept observations $p(z_{1:T}|y_{1:T})$ through a cascade of local importance distributions $p(z_{1:t}|y_{1:t})$. At each time t, p is the distribution over clusterings $z_{1:t}$ of observed concepts $y_{1:t}$, represented through the current set of particles. In order to compute the exact posterior distribution, the cat-



1	tree	apple	grow
2	sit	bird	tree
3	sweet	apple	eat
4	eat	pear	sweet
5	sit	dog	lawn

(a)
t=4
sample weights
t=5
resampling
rejuvenation
t=5 (revised)
(b)
p(k)
p(target | k)
p(context | k)
tree grow sit tree bird apple k=1
sweet eat eat sweet pear apple k=2
lawn sit sit tree bird dog k=1
sweet eat eat sweet pear apple k=2
k=3 grow tree apple

Figure 4.2: (a) Visualization of the particle filtering procedure in the BayesCat model using an example of a 3-particle filter. Each particle corresponds to a clustering of the observed stimuli up to time t (left), and the collection of weighted particles serves as the current approximation of the posterior distribution over clusterings (right). The 5 concepts observed by the filter are shown in the tables. We show one update step for all particles with stimulus 5, and one subsequent resampling and rejuvenation step. In the resampling step the highest-weight (red) particle is duplicated, replacing the lowest-weight (green) particle. In the rejuvenation step each particle revisits one previous categorization decision in light of all available evidence (e.g., the blue particle removes *apple* (from stimulus 1) from the {*bird, dog*} cluster); (b) a zoom into the blue particle at time t=4 (left) and time t=5 after rejuvenation (right). Each particle consists of a distribution over categories, and category-specific distributions over target types and over context types.

Algorithm 3 The particle filter for the BayesCat model.

```
Initialize particles by randomly partitioning first $d$ stimuli    ▷ Initialization
Initialize weights $\mathbf{w}^d = \frac{1}{N}$
for stimulus $t = [d+1\dots T]$ do
    for particle $n = [1\dots N]$ do
        $z_n^t \quad \sim q(z_n^{1:t-1}|y^{1:t-1})q(z_n^t|z_n^{t-1}, y^t)$    ▷ Particle Update
              $= p(z^t = i|z^{-t}, W_{-d}; \alpha, \beta, \gamma)$    Equation (4.5)
        $S_n^t \leftarrow (S_n^{t-1}, z_n^t)$
        $\tilde{w}_n^t = w_n^{t-1} \times P(y^t|z^{t-1})$    ▷ Weight Update
              $= w_n^{t-1} \times \sum_i p(z^t = i|z^{-t}, W_{-d}; \alpha, \beta, \gamma)$
    $\mathbf{w}^t \leftarrow normalize(\tilde{\mathbf{w}}^t)$
    if $ESS(\mathbf{w}^t) \leq thresh$ then    ▷ Resampling
        $\mathcal{P}(i) \leftarrow \{Mult(\mathbf{w}^t)\}_{i=1}^N$
        $\mathbf{w}^t = \frac{1}{N}$
        for particle n $\in \mathcal{P}(i)$ do    ▷ Rejuvenation
            for rejuvenation point $o = [1\dots O]$ do
                $d^o \sim uniform(1\dots t)$
                $z_n^{d^o} \sim P(z_n^{d^o}|z_{n\setminus -d^o}^t, y_t)$    Equation (4.5)
```

egorization of observations $y_{1:t-1}$ would need to be re-computed for each time step considering all observed evidence. The exact posterior distribution is, however, not incremental, because the computation time of the re-estimation of the density over all previous category assignments is not constant in the number of observed concepts. It is not tractable to sample from the local target distribution, and not cognitively plausible either since it assumes re-organization of semantic knowledge with every new observation. Figure 4.2a displays the estimation of the posterior density through weighted particles (indicated by the size of the circles) on the right-hand side; the current state of the corresponding particles is shown on the left-hand side.

Following the importance sampling framework, we choose a proposal distribution $q(\cdot)$ with which we can approximate the local target distribution more efficiently, and which has a constant computation time with respect to the number of observed concepts. In particular, we assume that once a concept has been assigned a category, this category is fixed:

$$\begin{aligned} q(z_{1:t}|y_{1:t}) &= q(z_1|y_1)\prod_{k=2}^{t} q(z_k|z_{1:k-1},y_{1:k}) \\ &= q(z_{1:t-1}|y_{1:t-1})q(z_t|z_{1:t-1},y_{1:t}) \\ &= q(z_{1:t-1}|y_{1:t-1})q(z_t|z_{t-1},y_t), \end{aligned} \tag{4.6}$$

Importantly, this distribution depends *only* on the label assignments in the previous time step z_{t-1} since all previous category assignments are fixed and encoded in this state. This process corresponds to lines 5–6 in Algorithm 3. In the final line of equation 4.6, the first term corresponds to the distribution over clusterings of the first $t-1$ observations, as represented by the current set of particles (i.e., the result of the previous iteration). The second term denotes the probability distribution over categories for the current input y_t, i.e., over all different ways in which the concept can be integrated into the current samples. We compute this distribution individually for each particle, sample its category from this distribution, and update the particle state with the new information. Figure 4.2a illustrates how each particle is updated individually after observing input stimulus 5.

The remaining question is the definition of the distribution over categories for the new observation. Importance sampling affords flexibility in selecting the proposal distribution $q_t(z_t|z_{t-1},y_t)$. We sample category z_t for the current concept observation y_t from its posterior distribution over categories:

$$q_t(z_t|z_{t-1},y_t) = p(z_t|z_{1:t-1})p(y_t|z_t), \tag{4.7}$$

taking into account prior information about category probability and the features of the observed concept. We finally weigh each sample n by its importance weight w_n which can be shown to correspond to the predictive likelihood of the current stimulus y_t, and the weights are normalized to sum to one (see lines 7–8 in Algorithm 3). A more detailed explanation of particle filtering can be found in the technical background Section 3.3.3.

Resampling By repeatedly sampling from local approximations to the target density, inaccuracies will inevitably accumulate. This phenomenon, called degeneracy, is a common problem with particle filters, and manifests in highly varying particle weights. For our model this means that many particles, or sampled categorizations, will not be representative of the categories present in the data. Ideally, however, a learner should focus on "good" hypotheses in order to use its capacities effectively. The "goodness" of a sample is indicated by its importance weight. A common approach to counteract accumulating errors, called *resampling*, is to replace low-weight particles with copies of high-weight particles based on some pre-determined schedule (see Section 3.3.3.3 for more information.). This way, memory resources can be allocated on high-probability particles, individual copies of which can be further propagated. We incorporate a threshold-based resampling scheme, using the *effective sample size* (ESS):

$$ESS(w^t) = \left(\frac{1}{\sum_n (w_n^t)^2} \right), \tag{4.8}$$

which is inversely correlated with the variance of the current set of particle weights. A resampling step is executed whenever the ESS falls below a set threshold. This threshold-based resampling provides a means of modeling memory limitations based purely on the learner's internal state. From a modeling perspective, this provides us with a statistically sound learning procedure, which is defined purely with respect to the current state of "confidence" of the learner, without the need to resort to external cues or heuristics. Figure 4.2a shows one resampling step following the particle updates. The red particle with the highest weight is duplicated and replaces the green particle with the lowest weight (see the different-sized circles on the right-hand side).

Technically, resampling consists of drawing N times with replacement from a multinomial distribution over particles parametrized by the current set of particle weights. Weights are re-set to uniform after resampling (see lines 9–11 in Figure 3). The resulting set of particles is an empirical estimate of the current approximation, in that the weights are now implicitly represented in the number of instantiations of the sampled particles. We use systematic sampling (Cochran, 1977) to obtain a new set of particles from the multinomial distribution, which has been shown to produce samples with less variance than simple multinomial sampling (Hol et al., 2006).

Relaxing Strict Incrementality The learning algorithm presented above approximates the target distribution over categorizations of observed concepts in a *strictly* incremental way. In other words, while it simulates human memory restrictions and uncertainty by learning based on a limited number of current knowledge states, it *never* reconsiders past categorization decisions. However, in many linguistic tasks, learners revisit past decisions (Frazier and Rayner, 1982) and intuitively we would expect categories to change based on novel evidence, especially in the early learning phase (Colunga and Smith, 2005; Landau et al., 1998; Borovsky and Elman, 2006). Children clearly revise and refine their early hypotheses of the world in light of new information.

We incorporate this intuition into our particle filter, by allowing it to reconsider past decisions to some extent, while keeping the algorithm incremental and computation time constant. We employ a technique called *rejuvenation* (Gilks and Berzuini, 2001). Specifically, after the resampling step for each particle, we individually reconsider the category assignment for a fixed number of previously observed concepts (see lines 13–15 in Figure 3). Aside from being cognitively plausible, rejuvenation also brings a theoretical advantage: it enhances the representativeness of the sample, by "jiggling" the resampled particles and thus introduces diversity among descendants of the same particle. Figure 4.2a illustrates rejuvenation for the bottom set of particles. Each particle revisits one previous categorization decision (e.g., the blue particle, places concept observation 1 into a previously empty cluster). Note that the previously identical copies of the red particle contain distinct clusterings after rejuvenation, such that the sample space is explored more effectively. See Section 3.3.3.3 for more information.

4.4 Experiment 1: Large-Scale Category Learning

In the following we present a series of experiments assessing the performance of BayesCat. Our experiments are designed to examine whether the model produces meaningful categories but also to investigate the learning process itself and its characteristics. In the first experiment (Section 4.4.1) we assess the quality of the semantic categories induced by our model and compare it against an ideal batch learner and Fountain and Lapata's (2011) incremental graph-based model. We continue with two experiments which explore category acquisition in children using a corpus of child-directed speech (Sections 4.5.1–4.5.2). Finally, Section 4.5.3 presents a typicality rating experiment. All our experiments evaluate the categories produced by the models against gold standard categories created by humans.

4.4.1 Quality of Learnt Categories

Our first goal was to examine whether any meaningful categories emerge when our incremental model is trained on a large corpus. We compare BayesCat against a related graph-based incremental learner, and a batch learning version of our own model. All models are trained on the British National Corpus (BNC), a 100 million word collection of samples of written and spoken British English.[4] Each model's resulting clustering is compared against a human-produced gold standard. In the following, we describe how this gold standard was created, discuss how model parameters were estimated and explain how model output was evaluated.

Data Our model was evaluated based on its clustering of words into semantic categories and its output was compared against similar clusters elicited from human participants. A gold standard set of categories was created by collating the resources developed by Fountain and Lapata (2010) and Vinson and Vigliocco (2008). Both data sets contain a classification of (concrete) nouns into (possibly multiple) semantic categories produced by human participants. Examples from the data set are provided in Table 4.2. The former data set is an extension of McRae et al.'s (2005) feature norms with category information. The original feature norms were collected through a major effort spanning multiple years and involving more than 700 participants. Norms were collected for a set of 541 target concepts consisting of living (e.g., *cow*) and non-living (e.g., *blender*) things, each corresponding to a single English noun. Concepts were selected so as to cover a broad range of generally familiar basic level concepts used in previous studies on semantic memory.

Fountain and Lapata (2010) augmented McRae et al.'s (2005) concepts with category labels (and typicality ratings). They collected this information using Amazon Mechanical Turk, an online labor marketplace which has been used in a wide variety of elicitation studies and has been shown to be an inexpensive, fast, and (reasonably) reliable source of non-expert annotation for simple tasks (Snow et al., 2008). Participants were presented with 20 randomly selected concepts from the McRae data set, and asked to write down the superordinate category they thought applied (rather than select one from a list). Each concept was labeled by ten participants. Based on the set of collected labels, the concepts were grouped into 41 categories (allowing for multi-

[4] The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

BUILDING
church, garage, skyscraper, tent, shack, wall, door, basement, house, pyramid, brick, cathedral, chapel, hut, apartment, cabin, bungalow, stone, barn

VEHICLE
yacht, unicycle, boat, raft, bus, train, bike, trailer, submarine, sled, truck, rocket, jet, van, subway, tractor, skateboard, trolley, helicopter, buggy, jeep, motorcycle, ship, canoe, ambulance, sailboat, airplane, limousine, sleigh, taxi, car, scooter, tank.

WEAPON
cannon, gun, machete, rifle, bayonet, harpoon, bazooka, tomahawk, whip, catapult, sword, revolver, knife, missile, bow, crowbar, shotgun, dagger, tank

Table 4.2: Example categories and their concepts taken from our gold standard.

category membership). The reliability of the annotations was assessed through labeling correlation between random splits of the data, and amounts to an average of 0.72 across all categories (ranging from 0.91 (FURNITURE) to 0.13 (STRUCTURE)). Given the elicitation procedure described above, we assume that the feature norms represent psychologically salient categories which the cognitive system is in principle capable of acquiring.

In order to evaluate category acquisition models on a large scale, we further merged McRae et al.'s (2005) data set with the concepts used in Vinson and Vigliocco (2008). The latter data set covers concrete basic level objects, event-related objects, and verbs, however in this work we only used the subset of 169 concrete objects. Category labels for these objects are provided by the authors and largely overlap with those elicited in Fountain and Lapata (2010). For this reason, we did not elicit additional category labels empirically. After removing duplicates, we obtained 42 semantic categories for 555 nouns. We split this gold standard into a development (70%; 41 categories, 492 nouns) and a test set (30%; 16 categories, 196 nouns).[5] The size and nature of this evaluation data set is in sharp contrast to those used in previous categorization studies which consist of a small number of artificial concepts.

The input to all models comprises the same set of linguistic stimuli, each of which

[5] The data set is available from `www.frermann.de/data`.

		BNC	CHILDES
Stimuli		1.37M	170K
Concepts	(target word types)	555	312
Features	(context word types)	6,584	2,756

Table 4.3: Number of stimuli, target concepts, and features retrieved from BNC and CHILDES.

consists of one target word t, surrounded by a symmetric window of n context words $[c_{-n} \dots c_{-1} \;\; t \;\; c_1 \dots c_n]$. The target words are defined by the set of concepts included in our gold standard. Some corpus statistics are given in Table 4.3 (column BNC). The corpus was lemmatized and stopwords were removed. Infrequent context words (occurring less than 800 times) were also eliminated, which leads to a reduction of context word types from 306,746 to 6,584 (by close to 98%) [6]. We used a window of size $n = 5$ for stimuli extracted from the BNC.

Model Comparison We optimized the parameters of the incremental BayesCat model on the development set. We obtained best results with the following parameters $\alpha = 0.7, \beta = 0.1, \gamma = 0.1$. Our model is parametric in the sense that the form of the model distributions are fixed to be K-multinomial. We set the maximum number of categories our model can learn to $K = 100$. However, the number of categories present in the data is much smaller, and the model reliably converges to using a subset of the 100 categories. For learning, we use a particle filter with $N = 100$ particles. We set the ESS threshold to $0.5 * N = 50$. After each resampling step we rejuvenate 100 randomly chosen previous categorization decisions, independently in each resampled particle.

We compare our BayesCat model against Fountain and Lapata's (2011) incremental model which adopts a graph-based approach to category learning. Concepts are represented as vertices in a graph and categories are inferred by grouping together distributionally similar vertices. The graph is partitioned into categories using an incremental variant of Chinese Whispers (Biemann, 2006), a non-parametric clustering algorithm (henceforth we refer to this model as CW). Their model implements category learning in the following steps. First, a semantic space is learnt — concepts are

[6] This drastic reduction follows from the fact that word type frequency in natural language follows a power law distribution, in particular Zipf's law. This implies that relatively few word types occur with high frequency and a large number of word types occur with low frequency.

represented as high-dimensional vectors, where each component corresponds to some co-occurring contextual element. Next, an undirected weighted graph $G = (V, E, \phi)$ is constructed with vertices V, edges E, and edge weight function ϕ. Concepts are added to the graph as vertices. Then, for each possible pair of vertices (v_i, v_j), their vector similarity $\phi(v_i, v_j)$ is computed and if the weight exceeds a threshold, an undirected edge $e = (v_i, v_j)$ is added to the graph. Finally, the graph serves as input to CW which produces a hard clustering over the graph vertices. The algorithm iteratively assigns cluster labels to vertices by greedily choosing the most common label amongst the neighbors of the vertex being updated. During this process, CW adaptively determines an appropriate number of clusters to accommodate the data. Both the semantic space, and the resulting graph are constructed incrementally, using co-occurrence counts collected from sequentially encountered input. Following Fountain and Lapata (2011), we transform co-occurrence counts into positive PMI values, and encode edge weights in the graph as cosine similarity values. We trained the CW model on the same set of stimuli as the BayesCat model, extracted from the BNC using a ±5 context window centered around the target concept mention. Edge weights must exceed a certain threshold in order for any two vertices to be clustered together. We tuned this threshold experimentally on the development data and obtained best performance with $t = 5$. We used this value in all our experiments.

The CW model treats semantic category acquisition and semantic knowledge representation as two different processes, even though it seems unlikely that humans have separate mechanisms for learning the meaning of words and their categories. Moreover, in contrast to BayesCat which learns category-specific features together with the categories, CW does not provide a straightforward way of recovering category-specific features from the clustered graph. We compared the learning behavior as well as the output clusters produced by the two models.

We also compared our incremental model against a batch learner which observes all input data from the start, as described in Section 4.3.1 The batch model (henceforth Gibbs) differs from the incremental BayesCat model *only* in its learning strategy and can thus be viewed as an ideal learner: it has access to all the training data at any time and can revisit previous categorization decisions systematically. We compare our incremental learner against an ideal batch learner, in order to investigate whether different learning strategies influence the quality of the estimated categories. Our experiments used the same model parametrizations for Gibbs as for the incremental BayesCat model. We run the sampler for 200 iterations without burn-in or lag, and take the state at the final iteration as our sample.

Method BayesCat produces soft cluster assignments, however, CW returns a set of hard clusters. In order to compare the two models directly, we transform soft clusters into hard clusters by assigning each target concept w to its most likely category z:

$$cat(w) = \max_{z} P(w|z) \cdot P(z|w) \tag{4.9}$$

The output clusters of an unsupervised learner do not have a natural interpretation. Cluster evaluation in this case involves mapping the induced clusters to a gold standard and measuring to what extent the two clusterings (induced and gold) agree (Lang and Lapata, 2011). Purity (pu) measures the extent to which each induced category contains concepts that share the same gold category. Let G_j denote the set of concepts belonging to the j-th gold category and C_i the set of concepts belonging to the i-th cluster. Purity is calculated as the member overlap between an induced category and its mapped gold category. The scores are aggregated across all induced categories i, and normalized by the total number of category members N:

$$\text{pu} = \frac{1}{N} \sum_{i} \max_{j} |C_i \cap G_j| \tag{4.10}$$

Inversely, collocation (co) measures the extent to which *all* members of a gold category are present in an induced category. For each gold category we determine the induced category with the highest concept overlap and then compute the number of shared concepts. Overlap scores are aggregated over all gold categories j, and normalized by the total number of category members N:

$$\text{co} = \frac{1}{N} \sum_{j} \max_{i} |C_i \cap G_j| \tag{4.11}$$

Finally, the harmonic mean of purity and collocation can be used to report a single measure of clustering quality. If β is greater than 1, purity is weighted more strongly in the calculation, if β is less than 1, collocation is weighted more strongly:

$$\text{F}_\beta = \frac{(1+\beta) \cdot pu \cdot co}{(\beta \cdot pu) + co} \tag{4.12}$$

In addition to purity and collocation and their harmonic mean, we report results using a fuzzy variant of the well-known *V-Measure* (Utt et al., 2014; Rosenberg and

Hirschberg, 2007) which is more appropriate for evaluating model output against the soft gold standard clusters.[7] V-Measure (VM) is an information-theoretic measure, designed to be analogous to F-measure, in that it is defined as the weighted harmonic means of two values, *homogeneity* (VH, the precision analogue) and *completeness* (VC, the recall analogue):

$$\text{VH} = 1 - \frac{H(G|C)}{H(G)} \tag{4.13}$$

$$\text{VC} = 1 - \frac{H(C|G)}{H(C)} \tag{4.14}$$

$$\text{VM} = 1 - \frac{(1+\beta) \cdot VH \cdot VC}{(\beta \cdot VH) + VC} \tag{4.15}$$

where $H(\cdot)$ is the entropy function; $H(C|G)$ denotes the conditional entropy of C given G and quantifies the amount of additional information contained in C with respect to C. The various entropy values involve the estimation of the joint probability of induced class C and gold standard class G:

$$\hat{p}(C,G) = \frac{\mu(C \cap G)}{N} \tag{4.16}$$

The fuzzy V-Measure distributes the mass of any object which is member of more than one cluster equally over all its clusters. Then, $\mu(C \cap G)$ is the total mass of the objects in the data shared by C and G and N the total mass of the clustering. As a result, N will be equal to the total number of objects to be clustered, which is trivially the case when comparing hard clusterings (but not for soft clusterings when the mass distribution step of the fuzzy V-measure is omitted, as in standard V-measure). Fuzzy VM thus allows us to directly evaluate the output of our models against our soft gold standard clustering, avoiding biases through the normalization constant, as implied in the standard V-Measure.

Results Table 4.4 reports results on the performance of our incremental BayesCat model (PF), its batch version (Gibbs), and Chinese Whispers (CW), all trained on the BNC. We present results on the test set (16 categories, 196 nouns) and the larger development set (41 categories, 492 nouns). We quantify model performance using

[7] Some categories such as ANIMAL and FOOD, or FRUIT and FOOD naturally share concepts in our gold standard.

	Development Set					
	pu	co	$F_{0.5}$	VH	VC	VM
Random	0.24	0.05	0.13	0.39	0.30	0.34
PF	0.59	0.31	**0.50**	0.47	0.42	0.44
Gibbs	**0.63**	0.24	0.47	**0.51**	**0.43**	**0.47**
CW	0.35	**0.55**	0.37	0.18	0.32	0.23
	Test Set					
	pu	co	$F_{0.5}$	VH	VC	VM
Random	0.52	0.11	0.30	0.67	0.39	0.50
PF	0.69	0.42	**0.61**	0.68	**0.50**	0.58
Gibbs	**0.76**	0.28	0.57	**0.78**	**0.50**	**0.61**
CW	0.40	**0.55**	0.42	0.26	0.36	0.30

Table 4.4: Performance of the particle filter (PF), its Gibbs sampling variant (Gibbs), Chinese Whispers (CW), and a random baseline (Random) on the British National Corpus (BNC). Boldface highlights the best performing model under each evaluation metric.

purity (pu) collocation (co), and their harmonic mean (with β set to 0.5) as well as the fuzzy version of V-measure (VM) and its homogeneity (VH) and completeness (VC) components. All scores are averaged over 10 runs.

In addition, we report performance on the same metrics for a baseline which assigns concepts to K categories at random (Random). The score reflects average performance of 10 random categorizations. K is set to the same value for all systems. Note that the V-measure scores for the baseline are surprisingly high, often beating CW. V-Measure is known to favor sparse clusterings of few gold instances (e.g., Vlachos et al. (2009); Rosenberg and Hirschberg (2007)). Throughout our experiments we set K to a high value (higher than the known number of categories in the gold standard), and set hyperparameters such that models are encouraged to induce the number of categories by themselves as a subset of K. A random baseline will populate all K categories (there's no incentive to leave some unpopulated) and thus produce a large number of categories with relatively few members each. Thus, random categorizations in this setting are quite precise, but recall will be low. V-measure has been shown to score such sparse solutions favorably (even if they are rated worse by humans), and the effect becomes stronger with fewer gold standard instances relative to K as is the case in many of our experiments.[8]

Comparison of the two incremental models, namely PF and CW, shows that our model outperforms CW under most evaluation metrics both on the test and development set. Under the VM evaluation metric, PF consistently outperforms CW. Gibbs, the non-incremental model version of our model, performs best overall. This is not entirely surprising. When BayesCat learns in batch mode using a Gibbs sampler, it has access to the entire training data at any time and is able to systematically revise previous decisions. This puts the incremental variant at a disadvantage since the particle filter encounters the data piecemeal and only periodically resamples previously seen stimuli. Nevertheless, as shown in Table 4.4, PF's performance is close to Gibbs using VM. Although the general pattern of results is the same on the development and test sets, absolute scores for all systems are higher on the test set. This is expected, since the latter contains fewer categories with a smaller number of concepts and more accurate clusterings can be (on average) achieved more easily.

Table 4.5 shows example categories learnt by the incremental BayesCat model. Each induced category is characterized by a set of concepts (top), as well as a set of features representing different aspects of the meaning of the category (bottom). For example, *train*, *bus*, and *boat* are members of the category VEHICLE. Induced features for this category refer to users of vehicles (e.g., *passenger, driver*) and the actions they perform on them (e.g., *drive, ride, park, travel, arrive*) as well as locations where vehicles are found (e.g., *road, railway, station*). Another category the model discovers corresponds to BUILDING with members such as *house, cottage, skyscraper.* Some of the features relating to buildings also refer to their location (e.g., *city, street, village, north*), architectural style (e.g., *modern, ancient*), and material (e.g., *stone*).

In addition to the final categories produced by the models, we are interested in their learning behavior. Figures 4.3 and 4.4 show the learning curves for the two incremental models, PF and CW. The learning behavior of the CW algorithm does not resemble a steady learning curve. This can can be explained by the fact that categories are built based on co-occurrence counts of target- and context words. With an increasing number of observations, these counts become less distinctive between target concepts. Inspection of the output of the CW algorithm, reveals that it induces one very big cat-

[8] Technically this is the case because the distribution over gold classes within each induced cluster becomes peakier (precision gets better), while the distribution over induced clusters for each gold class does not change as drastically (recall decreases at a slower rate). This has been noted repeatedly (e.g., Rosenberg and Hirschberg (2007); Vlachos et al. (2009); Reichart and Rappoport (2009)).

BUILDING
house, building, wall, stone, bridge, cottage, gate, brick, inn,marble, hut, corn, pier, cellar, basement, canary, skyscraper, beehive
house, building, build, street, town, century, village, stone, garden, city, london, live, centre, modern, hall, family, site, design, ancient, north, tower, bridge, mill, museum

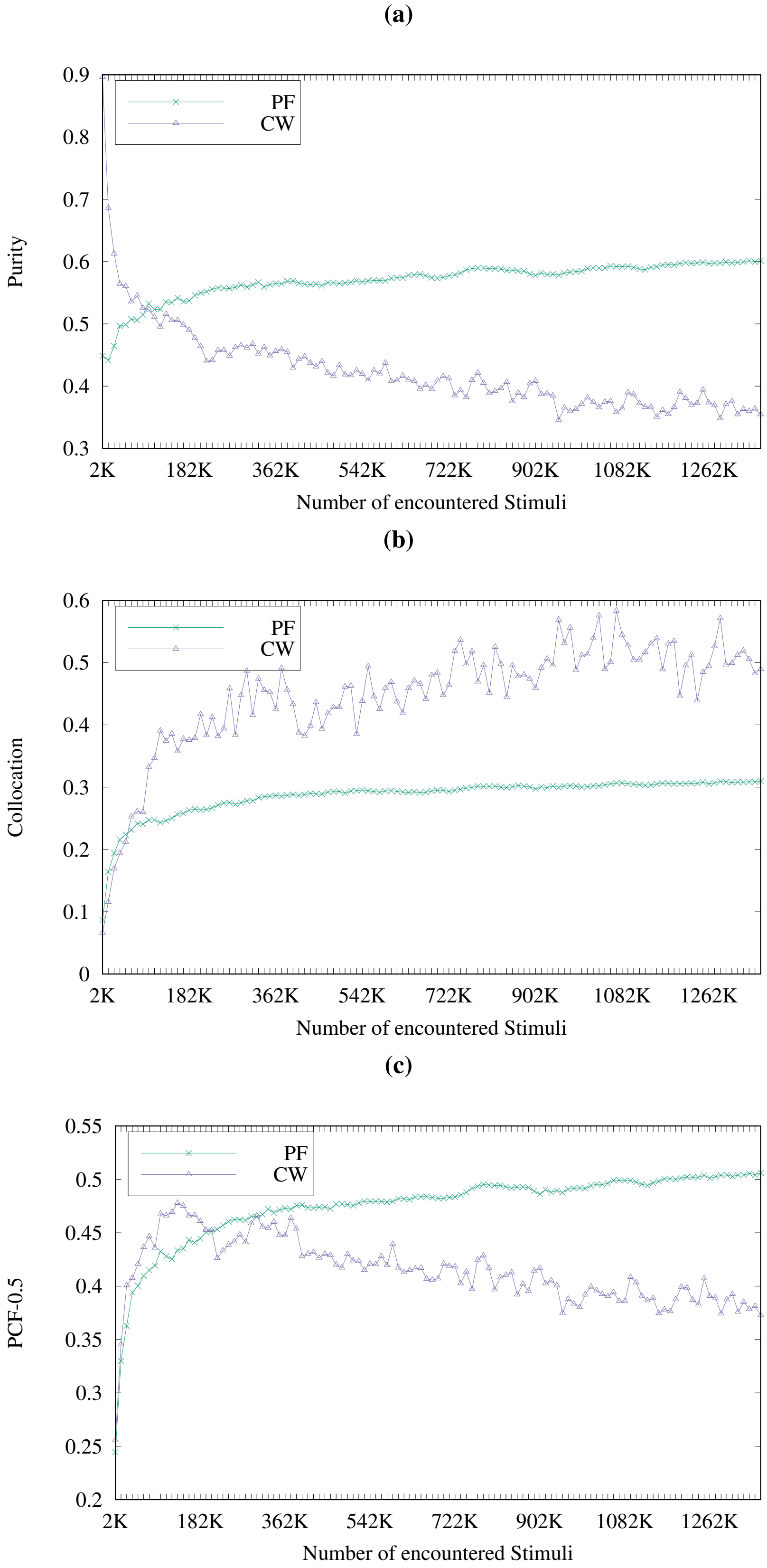
VEHICLE
train, bus, boat, wheel, van, truck, taxi, helicopter, garage, wagon, fence, bicycle, shed, trailer, cabin, tractor, cart, jeep, trolley, motorcycle, subway, escalator, airplane
car, road, drive, train, park, station, driver, bus, hour, line, fire, mile, vehicle, engine, passenger, boat, railway, travel, speed, arrive, track, traffic, route, yard, ride, steal

WEAPON
bomb, crown, knife, ambulance, bullet, shotgun, grenade, machete
police, court, home, hospital, die, kill, yesterday, attack, death, wife, injury, charge, officer, murder, shoot, suffer, arrest, victim, accident, parent, damage, injure, trial

INSTRUMENT
guitar, rock, piano, drum, violin, flute, clarinet, trumpet, cello, stereo, trombone, harp, harpsichord, rocker, accordion, saxophone, tuba, baton, bagpipe, harmonica
play, music, guitar, sound, band, bass, song, piano, instrument, sing, album, string, pop, drum, tune, violin, orchestra, dance, recording, solo, musical, performance, flute, mozart

Table 4.5: Examples of categories learnt from the BNC with the incremental BayesCat model. Category concepts (upper row) are shown together with their most likely features (lower row).

egory, comprising almost all of the target concepts, and a few rather small, but meaningful categories. On the contrary, the learning curves produced by the incremental BayesCat model show steady improvement of the acquired categories over time.



(a)

(b)

(c)

Figure 4.3: Learning curves for PF and CW on the BNC using (a) purity, (b) collocation, and (c) $F_{0.5}$.

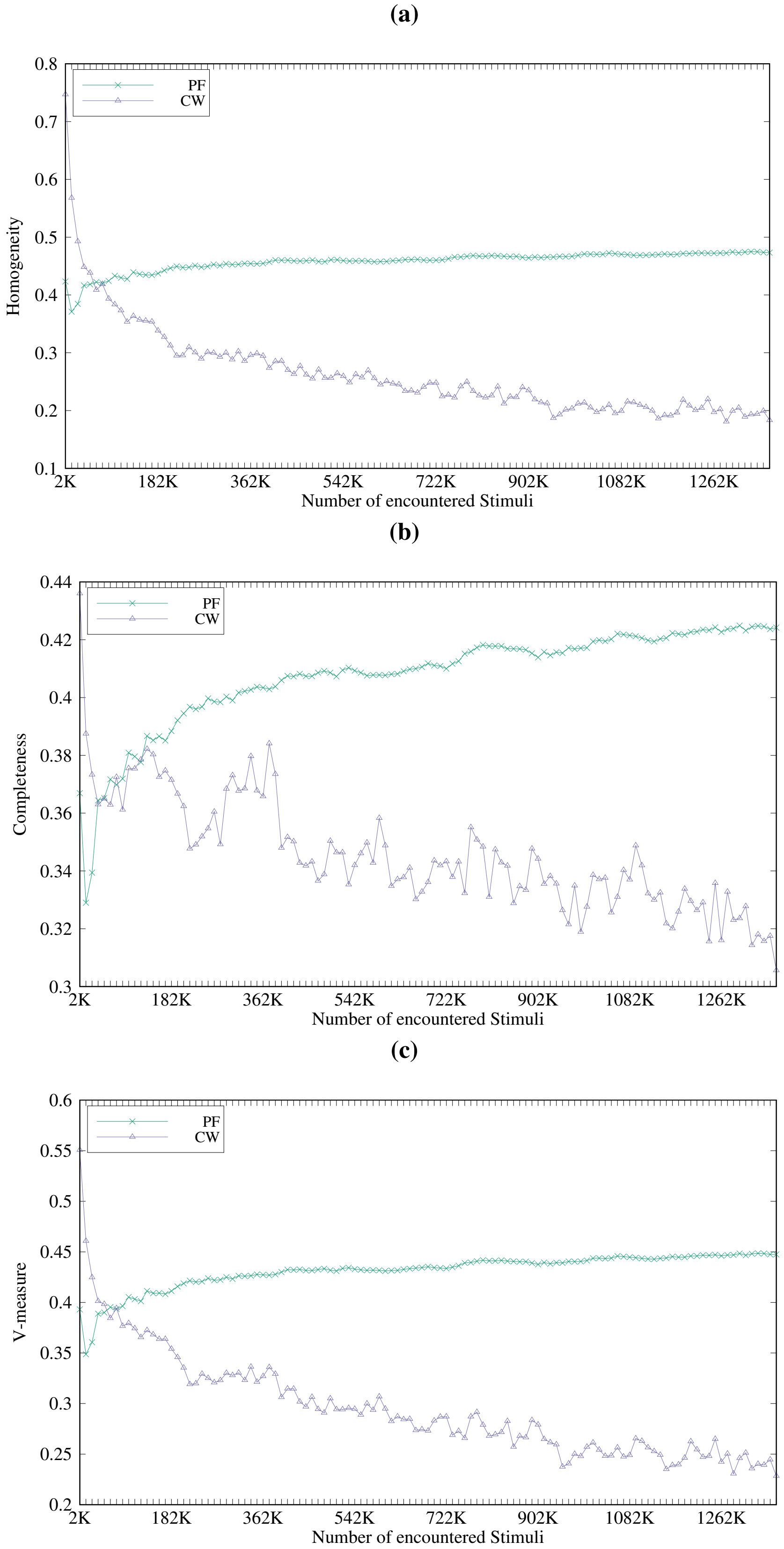


Figure 4.4: Learning curves for PF and CW on the BNC using (fuzzy) homogeneity (a), completeness (b), and V-measure (c).

4.4.2 Discussion

In this experiment, we performed a large-scale comparison among three models of natural language categorization. The incremental BayesCat model performs comparably to a batch version of the same model, showing a slightly worse performance. This seems to indicate that the Gibbs sampler provides a better fit to the cognitive gold standard and is to be preferred over the incremental learner. The learning process of the Gibbs sampler is, however, not cognitively plausible. While the latter is an ideal learner, with access to all data points at any time, and the ability to revise decisions systematically, it does not have a significant advantage over our incremental model. The Gibbs sampler can explore the search space more exhaustively than the incremental learner and can draw more accurate conclusions. Incremental learning highly depends on sufficient training data, and one would anticipate the particle filter's performance to increase with more observations.

Overall, the competitive performance of the particle filter is an encouraging result underlining the efficiency of the incremental learning paradigm as a basic characteristic of human cognitive behavior. Previous work (Fearnhead, 2004) has shown that Particle Filters outperform Gibbs samplers in Bayesian mixture models similar to the one presented here. Intuitively, the particle filter estimates a distribution over categorizations by means of its $N \geq 1$ incrementally constructed particles, or samples, which explore the probability space independently and simultaneously. A Gibbs sampler produces samples from a distribution by moving between different (high-probability) regions. This can be a very slow process, especially with many hidden variables involved, so that in practice a point estimate of the posterior distribution is often obtained.

We furthermore showed that the Bayesian models outperform a graph-based model of category acquisition. The categorizations learnt by CW reliably consist of one big category, comprising the vast majority of concepts, and very few small categories. The reported collocation and $F_{0.5}$ scores for CW are therefore misleadingly high: one large category results in a very high collocation score, while cluster purity remains very low throughout (see Figure 4.3a). For the incremental BayesCat model, however, the purity of categories improves constantly as well as well their completeness (see Figures 4.3a and 4.3b). The fuzzy V-measure does not overestimate CW's completeness score, and thus lends itself as a more suitable evaluation metric (see Figure 4.4).

In addition to its superior performance, we argue that BayesCat is also more cognitively plausible compared to CW. Firstly, on account of its architecture all information is represented in the same space as probability distributions over words and categories. In contrast, CW represents information as a co-occurrence matrix which needs to be transformed into a graph in order to learn categories. Secondly, the BayesCat model naturally induces category features during the process of category learning. Since features have been established as a good proxy for category representations in human cognition, it is inevitable that these representations evolve and change jointly while forming categories. CW only considers features in its first representation, the co-occurrence matrix, and there is no natural way of recovering category-specific features from the graph after categories have been learnt. From a cognitive point of view this separation is implausible. Experimental studies show that category and feature learning mutually influence each other (Goldstone et al., 2001; Schyns and Rodet, 1997): concepts are categorized based on their features, and the perception of features is influenced by already established categories. Like categories, features also evolve over time.

4.5 Experiment 2: Child Category Acquisition

The primary goal of the preceding experiment was to explore how effectively our model captures large-scale category information. Of equal interest, however, is modeling children's performance on an acquisition task — determining whether the linguistic input to which children are exposed enables learning of high-level semantic categories such as those seen in experiment 1. To answer this question, we applied our incremental model to a corpus of child-directed language and evaluated the resulting categories against the gold standard clusters used previously.

4.5.1 Quality of Learnt Categories

Data The CHILDES corpus (MacWhinney, 2000) was used to construct training stimuli for our model. CHILDES consists of a large number of transcripts in a multitude of languages, each recording a free-form interactive session between a child and one or more adults (parents); we used the XML portion of the corpus, consisting of American and British English transcripts.[9] All child produced utterances were excluded from the final set of stimuli. We extracted 170,000 child-directed stimuli which

[9] http://childes.psy.cmu.edu/data-xml/.

	pu	co	$F_{0.5}$	VH	VC	VM
Random	0.29	0.05	0.16	0.46	0.35	0.40
PF	0.62	0.21	0.45	0.50	0.42	0.45
Gibbs	**0.74**	0.19	**0.47**	**0.59**	**0.46**	**0.51**
CW	0.39	**0.54**	0.41	0.22	0.37	0.27

Table 4.6: Performance of the Particle Filter (PF), its Gibbs-based variant (Gibbs), incremental Chinese Whispers (CW), and a random baseline (Random) on the CHILDES corpus. Boldface highlights the best performing model under each evaluation metric.

we grouped according to the age of the child the speech was directed at.[10] The data was presented to the models in chronological order. Details about the size of CHILDES are provided in Table 4.3.

The corpus was lemmatized and stopwords were removed. Some concepts in the gold standard are very specialized and occur very infrequently or not at all in CHILDES. We only extracted stimuli containing target concepts occurring 50 times or more within the corpus. Analogously, we filtered low-frequency context words with the same threshold (leading to a reduction of context word types from 24,008 to 2,756 (by 89%)). Compared to the models trained on the BNC, we used a smaller context window size of $n = 2$. Child-directed utterances in CHILDES are relatively short and thus a small context window is necessary to capture linguistic features relevant to the meaning of the target concept.

The hyper-parameters of the BayesCat model were optimized on the BNC corpus (development set). We did not re-tune model parameters for CHILDES, and thus used the entire gold standard for evaluation (42 categories, 312 concepts). Model performance was assessed similarly to Simulation 1 using purity, collocation and their harmonic mean as well as the analogous information theoretic measures of homogeneity, completeness, and V-measure.

Results Table 4.6 presents our results on the CHILDES corpus. Again, we compare our incremental BayesCat model using a particle filter (PF), a batch version of the same model (Gibbs), incremental Chinese Whispers (CW), and a random baseline (Random). Please refer to Section 4.4.1 (page 86) for an explanation of the baseline

[10] Stimuli were binned in intervals of six months.

CLOTHES
hat, shirt, dress, pant, trouser, slipper, coat, suit, vest, jacket, glove, scarf, bow, tie
hat, wear, shirt, blue, daddy, color, dress, yellow, pant, slipper, coat, vest, got, scarf, short, button, clothes, bow, change, glove, cold, lovely, pretty, party, warm, suit, pocket

BODY PARTS
head, eye, nose, mouth, leg, tongue, chin, lip, shoulder
your, my, eye, nose, head, mouth, hurt, bump, pull, bite, blow, funny, silly, kiss, careful, tongue, chin, sore, ah, tickle, hard, touch, hole, fell, cry, matter, tire, body, shoulder

FRUIT
apple, cup, orange, strawberry, pear, plum, grape, banana, peach, saucer, lemon, raspberry, mug
eat, apple, hungry, cup, pear, orange, strawberry, grape, banana, green, wednesday, thursday, tuesday, fruit, plum, peach, monday, friday, peel, saucer, lemon, saturday

VEHICLE
car, train, truck, bridge, ambulance, van, tractor, crane, garage, trailer, taxi
car, oh, train, truck, thomas, drive, red, police, driver, engine, track, bridge, race, happen, people, ambulance, choo, park, road, station, mean, digger, saw, carry, trailer

Table 4.7: Examples of categories learnt from the CHILDES corpus with the incremental BayesCat model. Category concepts (upper row) are shown together with their most likely features (lower row).

and its misleadingly high V-measure scores. All scores are averaged over 10 runs. The results are broadly comparable to those obtained from the BNC. Again, we observe that Gibbs performs overall best, however, the incremental model is only slightly less accurate while being more cognitively plausible. Our model outperforms CW under most evaluation metrics. Examples of the semantic categories induced by BayesCat are shown in Table 4.7.

Figures 4.5 and 4.6 show how the clusterings evolve over time for the two incremental models (PF and CW). Again, CW does not show a meaningful learning curve, under any measure. The completeness of clusters increases over time, however, at the

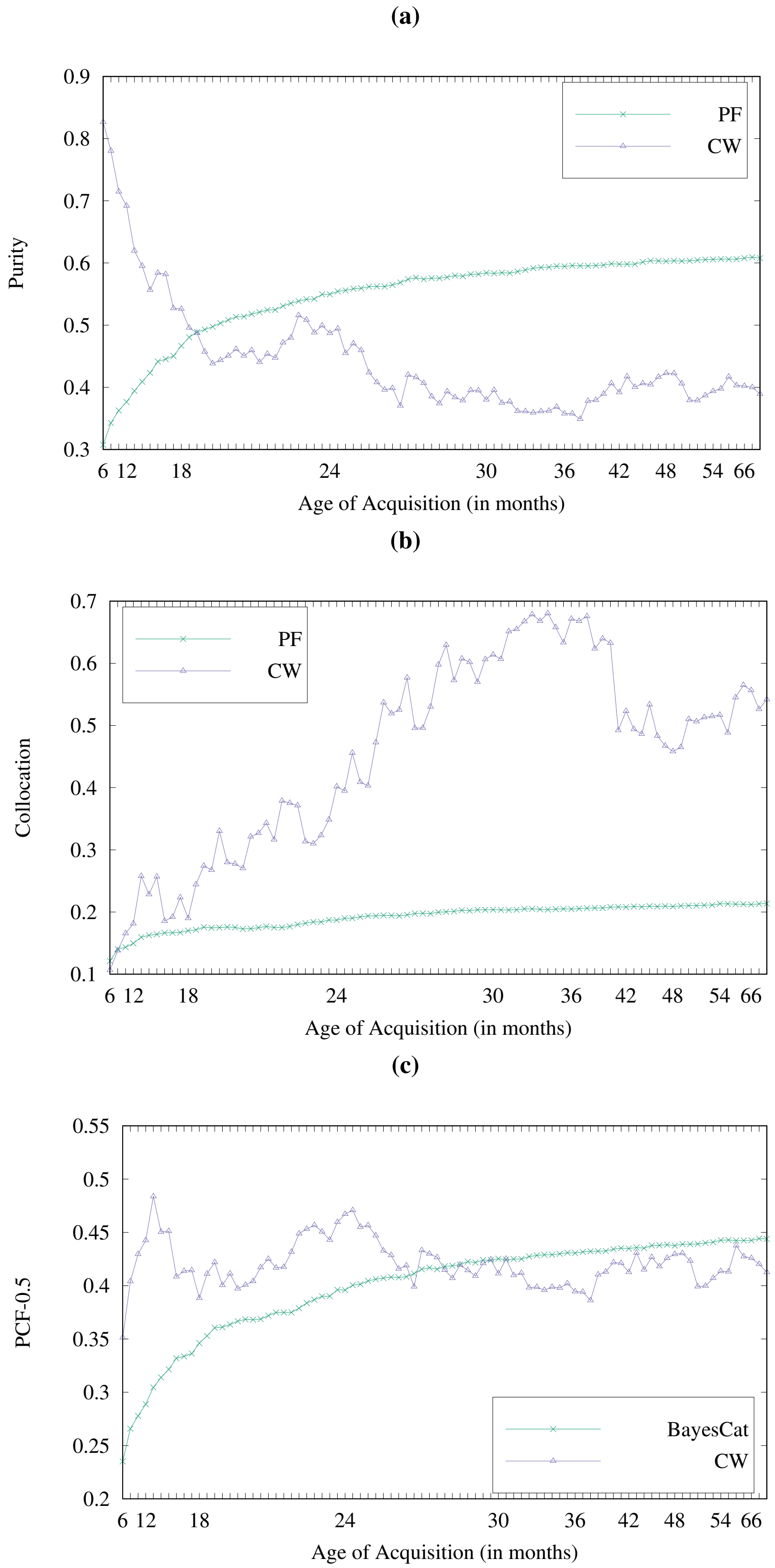
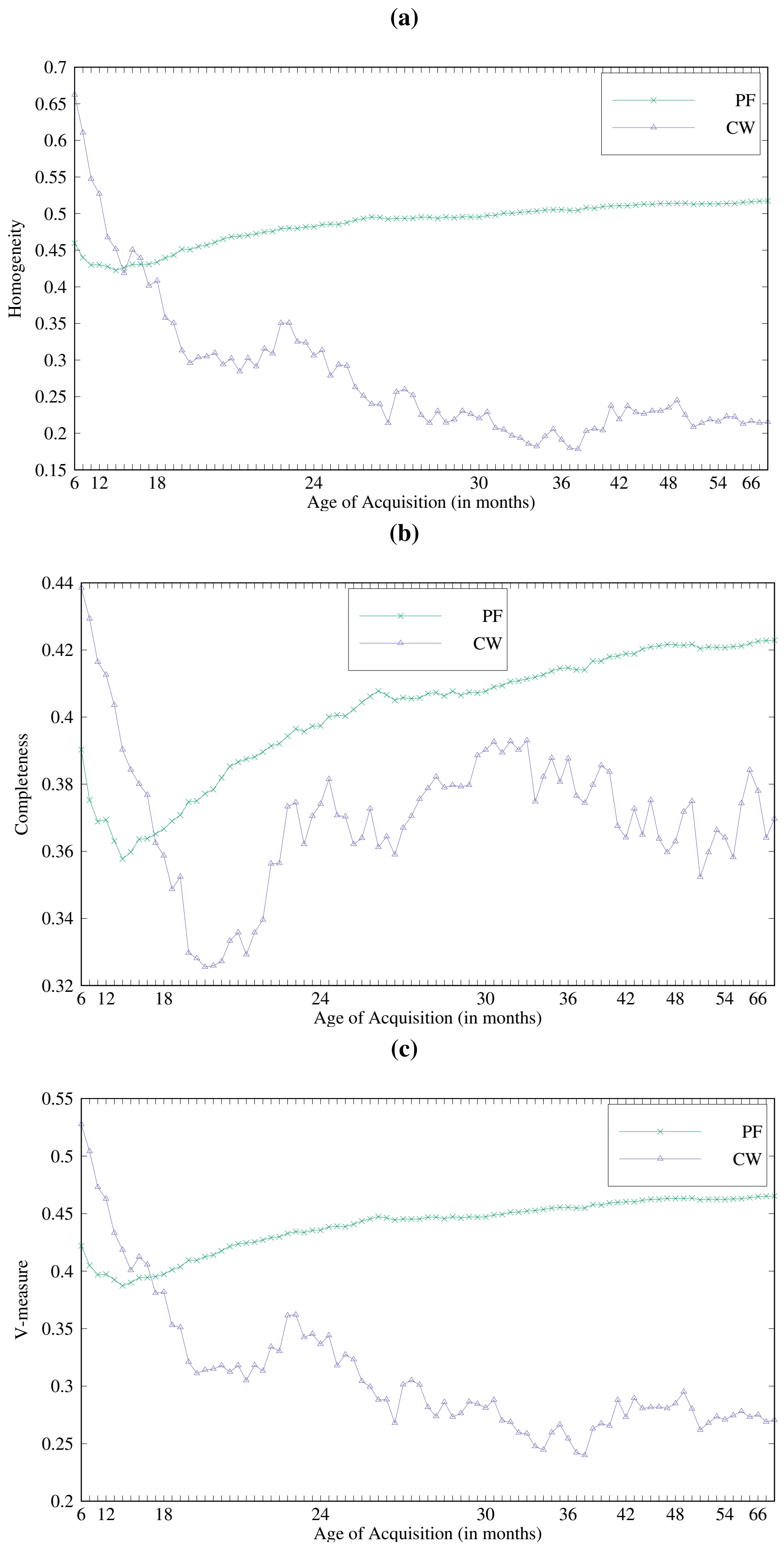


Figure 4.5: Learning curves for PF and CW on the CHILDES corpus using (a) purity, (b) collocation, and (c) $F_{0.5}$.



(a)

(b)

(c)

Figure 4.6: Learning curves for PF and CW on the CHILDES corpus using (fuzzy) (a) homogeneity, (b) completeness, (c) and V-measure.

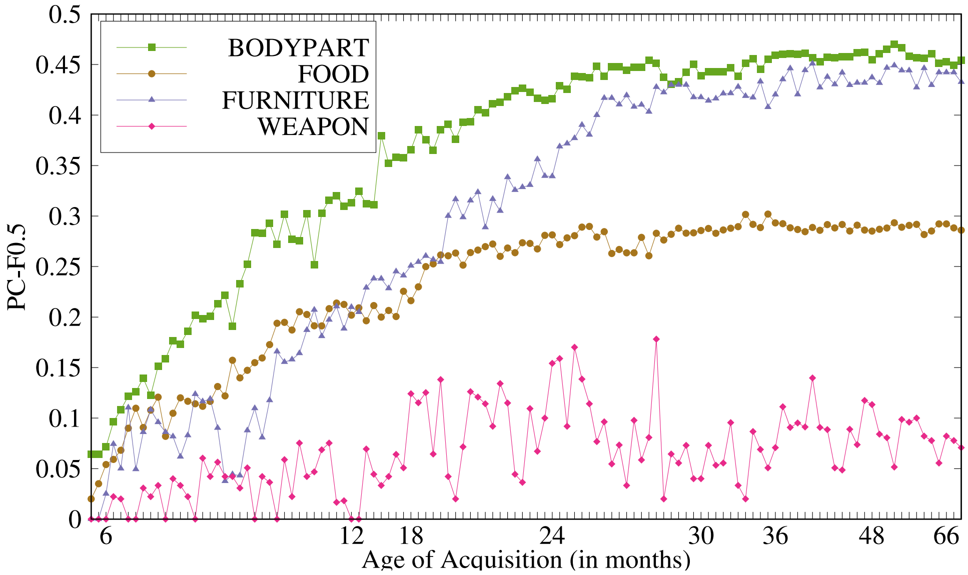


Figure 4.7: Emergence of selected categories over time for the incremental BayesCat model on the CHILDES corpus.

expense of purity. This effectively means that CW tends to learn one very big cluster comprising of the majority of target concepts. PF, on the other hand, shows clear learning curves across metrics, with increasingly clean (Figures 4.5(a) and 4.6(a)) and complete clusters (Figures 4.5(b) and 4.6(b)).

In addition to our quantitative evaluation against a gold standard, we investigated the learning process more qualitatively by inspecting the emergence of individual categories over time. Figure 4.7 shows how the categories BODYPARTS, FOOD, FURNITURE, and WEAPON develop in the course of 66 months. We can see that the category BODYPARTS emerges earliest and is acquired with high quality. The same is true for the category CLOTHES (not shown in the figure to avoid clutter). Slightly later, the categories FOOD, VEHICLES (also not shown), and FURNITURE evolve. Categories like, WEAPONS, however, are not acquired from the CHILDES corpus, presumably because care takers rarely talk about or use concepts from this category in the presence of young children. In contrast, the WEAPONS category is acquired from the BNC (see Table 4.5), which, again, emphasizes the ability of our model to adapt to and learn from empirical data.

4.5.2 Analysis of Memory Constraints

In this experiment we delve deeper into our incremental inference algorithm and its appropriateness for cognitive learning. While humans are generally very successful learners, their memory and computing power is clearly constrained. Particle filters provide us with a flexible way for investigating memory constraints. The *number of particles*, or hypotheses, available to the filter during learning directly correlates with its memory usage. We expect that, while humans do not have the means to entertain an exceeding number of hypotheses at any time, constraining the learner to one hypothesis will have a negative impact on the learning outcome. A second indicator of memory usage is *rejuvenation*, the extent to which past categorization decisions are being re-considered in the light of new evidence. Rejuvenation in the BayesCat model is tightly coupled with *resampling*, replacing low-probability particles with high-probability ones, which is yet another an indicator of cognitive load. Resampling (and rejuvenation) is driven by a learner-internal state of "confidence", where the model state is re-considered whenever the learner falls below a confidence threshold about earlier categorization decisions in the light of new evidence. A learner's confidence w.r.t. to the learnt categorization should increase over time, so that revisions of the model state occur less frequently. To summarize, in this set of experiments, we investigate two questions: (1) How do the number of particles and the extent of rejuvenation influence the learning process and the quality of the learnt categorization; and (2) how does the extent of resampling evolve over time.

Method We compare particle filters with different numbers of particles n, where $n \in \{1, 5, 20, 50, 100\}$. The number of particles is the only varying experimental variable, and the particle filters are set up as described in the previous experiments. Resampling takes place if the ESS falls below a pre-specified threshold; rejuvenation (of 100 stimuli) occurs after every resampling step. For the sake of brevity, we present results on CHILDES only, noting that a very similar picture emerges on the BNC. The training corpus used in this set of experiments is identical to the one used in the category quality evaluation in the previous section.

We compare the performance of the particle filters using two different metrics. First, we report learning curves based on model log-likelihood. The log-likelihood is a common model-internal metric used for measuring convergence, even though it does not necessarily correlate with the usefulness or interpretability of the estimated solution (Chang et al., 2009). A higher log-likelihood indicates a better model. In order to directly measure the quality of the categorizations induced by the particle filters, we additionally report the $F_{0.5}$ measure. Moreover, we are interested in teasing apart how the number of particles and rejuvenation influence the learning behavior of our model. To this end, we compare particle filters with differing numbers of particles, but with rejuvenation disabled.

Results Figures 4.8a and 4.8b show the log-likelihood-based learning curve produced for particle filters with a varying number of particles. While the shape of the curve is very similar across particle filters, a substantial improvement from the one-particle filter to multiple-particle filters can be observed. However, the improvement decreases with more particles, although a slight advantage is still observable. A very similar picture emerges for the learning curves based on category quality (Figure 4.8c). The categorizations inferred by the one-particle filter are less accurate than those inferred by multiple-particle filters. This suggests that the one-particle filter found a local maximum, from which it could not escape. The advantage of the Gibbs sampler as an ideal learner becomes apparent with the log-likelihood metric (see the red point Figure 4.8a). The BayesCat model using Gibbs sampling achieves significantly better log-likelihood scores compared to the incremental model. In general, we see an initial improvement in the learning curve, but a subsequent drop which is caused by the increasing number of input stimuli which need to be integrated into a coherent categorization. The log-likelihood flattens out towards the end of the learning curve. While ideally it should eventually improve, we suspect that the size of the stimuli set used in this experiment was too small.

Figure 4.9 compares the learning curves for different particle filters with rejuvenation disabled. Across filters and evaluation metrics a clear decrease in performance is observed, which is unsurprising given that the filters now are bound to categorization decisions, and unable to revise past decisions in the light of new experience. It is still evident, however to a lesser extent, that the one-particle filter performs worse compared to filters with more than one particle. Especially in the early learning phase, the ability to explore multiple hypotheses in parallel is advantageous (see Figure 4.9b).

Figure 4.10 illustrates the *resampling* behavior of the particle filters. On the one hand, we observe that filters with more particles tend to resample more frequently, i.e., the

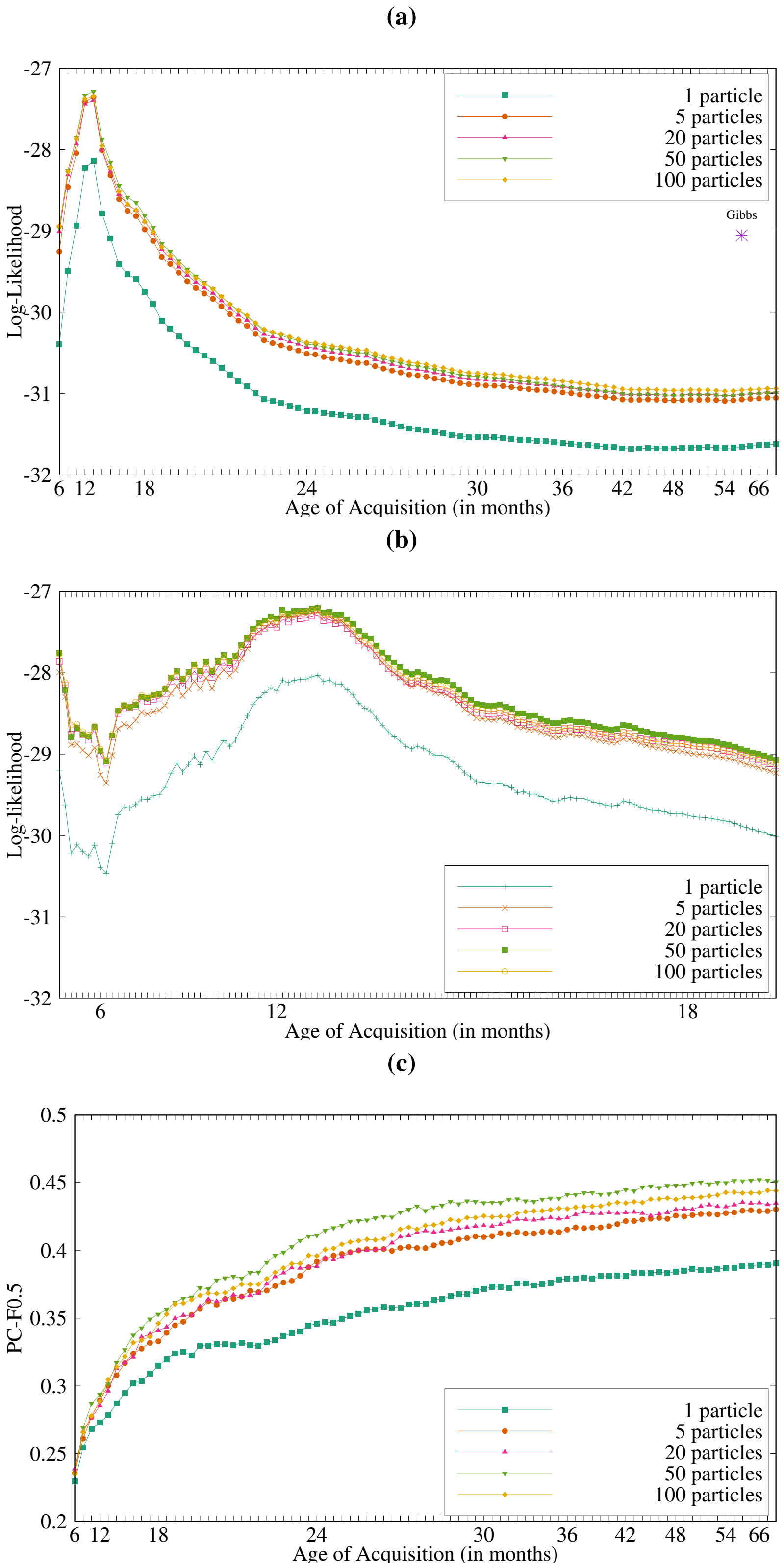
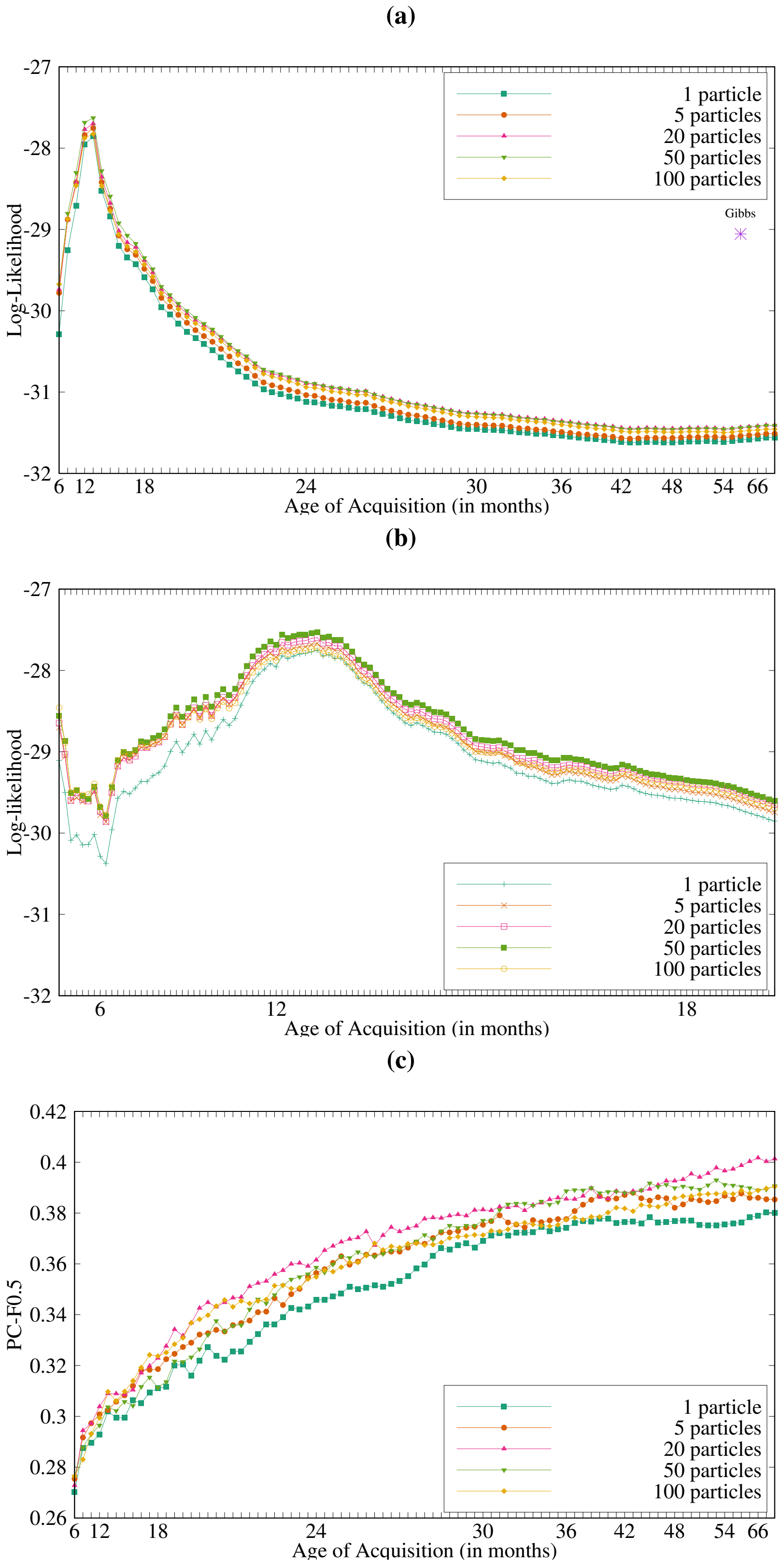


Figure 4.8: Learning curve for the BayesCat model on CHILDES with varying number of particles. Model log-likelihood curve (a), model log-likelihood curve for the early learning phase (b), and $F_{0.5}$ learning curve (c).


(a)

(b)

(c)

Figure 4.9: Learning curve for the BayesCat model on CHILDES with rejuvenation disabled. Model log-likelihood curve (a), model log-likelihood curve for the early learning phase (b), and $F_{0.5}$ learning curve (c).

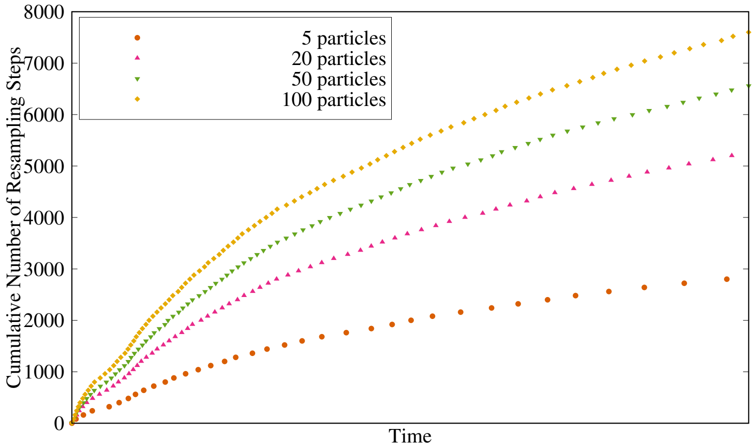


Figure 4.10: Resampling behavior of the BayesCat model learnt with a varying number of particles. Points correspond to executed resampling steps at time x.

weights of the particles tend to diverge more with an increasing number of particles. On the other hand, across different filters resampling frequency decreases over time, confirming our intuition that a learner's knowledge state should become increasingly confident over time, and reconsiderations of past decisions decrease in frequency.

4.5.3 Typicality Rating

An important finding in the study of natural language concepts is that categories show graded category-membership structure. For example, humans generally judge a *trout* to be a better example of the category FISH than *eel*. In the same way, an *apple* intuitively seems to be a better example of the category FRUIT than *olives*. Several experimental studies underline the pervasiveness of typicality (or "goodness of example") in a wide variety of cognitive tasks such as priming (Rosch, 1977), sentence verification (McCloskey and Clucksberg, 1979), and inductive reasoning (Rips, 1975). Because of its importance, typicality is also an evaluation criterion for models of categorization and concept representation. Any such model should be able to give an account of the graded category structure and correctly predict differences in the typicality of category members.

We therefore assessed our model on a typicality rating task (Voorspoels et al., 2008). In this task, the model is presented with concepts of a category and must predict the degree to which the concepts are typical amongst members of that category.

Method Previous work on semantic categorization has shown that exemplar models perform consistently better compared to prototypes across a broad range of linguistic tasks (Voorspoels et al., 2008; Fountain and Lapata, 2010; Storms et al., 2000). This finding is also in line with studies involving artificial stimuli (e.g., Nosofsky 1992). For the typicality rating task we therefore adopted an exemplar-based model which is broadly similar to the generalized context model (Nosofsky, 1984, 1986). In this model, a measure of the typicality of a concept is derived by summing the similarity of that concept to all concepts in the category. More formally, the typicality of concept w for category G is given by:

$$T_G(w) = \sum_{v \in G} \eta_{w,v} \tag{4.17}$$

where $\eta_{w,v}$ is the similarity of concept w to concept v, with v also belonging to category G. The similarity function $\eta_{w,v}$ can vary depending on how concepts and categories are represented (e.g., spatially or probabilistically). Within our Bayesian framework it is relatively straightforward to specify a probabilistic quantity that corresponds to the strength of association between w and v (Griffiths et al., 2007b):

$$\begin{aligned}\eta_{w,v} = P(v|w) &= \sum_k P(v|k)P(k|w) \\ &= \sum_k P(v|k)\frac{P(w|k)P(k)}{P(w)}\end{aligned} \tag{4.18}$$

Here the probability of a category given concept w and the probability of concept v given that category are averaged across all categories k.

In this set of experiments, we compared BayesCat against a simple co-occurrence based model, essentially identical to the semantic space used as input to CW. In this space each target concept is represented as a vector with dimensions corresponding to its co-occurring context elements. As in previous experiments, we transformed raw co-occurrence counts into PMI values. A typicality value for each member of a category was computed using (4.17) and summing the cosine similarity of the concept vector $\overrightarrow{w}$ to the all other vectors representing its co-members $\overrightarrow{v}$:

$$\eta_{w,v} = cos(\overrightarrow{w}, \overrightarrow{v}) = \frac{\overrightarrow{w} \cdot \overrightarrow{v}}{|\overrightarrow{w}||\overrightarrow{v}|} \tag{4.19}$$

Our experiments used the data set produced by Fountain and Lapata (2010) who elicited typicality ratings[11] (and category labels) for all concepts contained in the feature norms

[11] Publicly available from `http://homepages.inf.ed.ac.uk/s0897549/data/`.

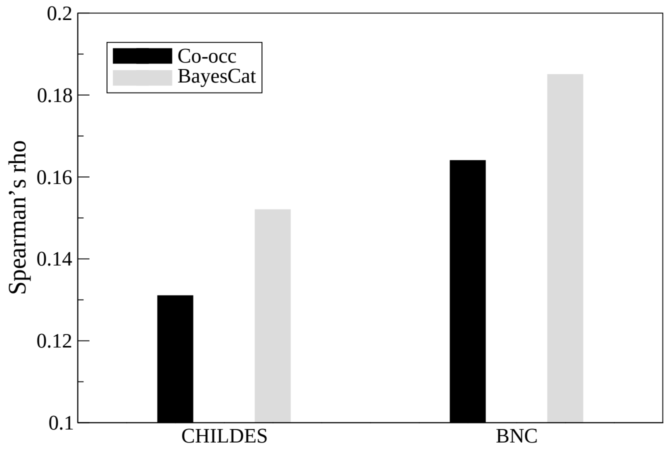


Figure 4.11: Rank correlations (Spearman's rho) between the gold typicality ranking and the model produced rankings over the set of all gold standard categories.

of McRae et al. (2005). In the evaluation, we present the models with the set of gold members of each gold category, and compare the rankings produced by the models with the gold typicality ranking elicited from humans. We report Spearman's ρ correlation co-efficients for the global ranking across all categories in this data set. We present results on the CHILDES corpus (41 categories, 689 concept-category pairs) and the BNC (41 categories, 1,226 concept-category pairs). Typicality ratings were produced with the incremental variant of the BayesCat model trained with 100 particles. Our results are averaged over 10 runs. The co-occurrence based model is deterministic, hence we only report one run for that model.

Results Our results are summarized in Figure 4.11 which illustrates model performance (as measured by Spearman's rho) on the BNC and CHILDES. The incremental BayesCat model is consistently better at predicting typicality ratings compared to the simpler co-occurrence based model. All correlation coefficients in Figure 4.11 are statistically significant ($p < 0.01$). We should also point out that the typicality rating task is generally difficult even for humans. Fountain and Lapata (2010) measured inter-subject agreement in their elicitation study to 0.64. BayesCat fits the experimental data better when trained on the BNC. This is not unexpected since the BNC is much larger than CHILDES by a factor of almost 10. Table 4.8 shows some qualitative examples of concepts which BayesCat rated as most typical/atypical for a particular category.

category	most typical concepts	least typical concepts
FOOD	*cake, bread*, strawberry, cheese*	*owl*, lobster, snail*, deer**
ANIMAL	*elephant, horse, cow*, duck*	*bat, pickle, chipmunk, tuna**
CLOTHING	*shirt*, shoe, sock, dress**	*necklace*, cap, cape, hose**
VEHICLE	*car, train*, truck*, bus**	*ship, tank, motorcycle, trolley*

category	most typical concepts	least typical concepts
FOOD	*cheese, bread*, cake, potato*	*honeydew, blueberry, eggplant, zucchini*
ANIMAL	*dog, bear, horse, cat**	*chipmunk*, chickadee, bluejay, groundhog*
CLOTHING	*dress*, shirt*, shoe, jacket*	*nightgown, mitten, earmuff, pajamas*
VEHICLE	*car, train*, bus*, ship*	*surfboard*, sled*, sleigh, unicycle*

Table 4.8: Qualitative examples of typicality judgments as predicted from the incremental BayesCat model trained on CHILDES (top) and the BNC (bottom). The four most typical concepts, and the four least typical concepts are displayed for selected categories. Superscript * indicates whether the concept was deemed highly typical/atypical in Fountain and Lapata's (2010) elicitation study.

4.5.4 Discussion

The preceding series of experiments investigated category learnablility from a corpus of child-directed language and showed that meaningful categories emerged from the BayesCat model. Compared to our large-scale experiments on the BNC, our model was presented with a smaller amount of stimuli, and yet was able to recover semantic categories without any corpus specific optimization. This highlights the robustness of our model with respect to the chosen hyper-parameters or training corpus. Note, however, that the runtime of the incremental filter is linear in the number of input stimuli, and thus is efficiently applicable to data sets of increasing size.

The qualitative examples of the categories and features learnt by BayesCat (Table 4.7) demonstrate that the categories and their associated features are coherent and easily interpretable. Note that concepts and features are not clearly separated: frequent members of a category also appear in its feature set. We do not treat concepts and their features differently. From a cognitive point of view this is plausible: concepts of the same category can be co-observed (e.g., one may wear a hat and coat or eat an apple and a banana) which seems like a useful signal in category learning.

Beyond the quality of learnt categories, our experiments also analyzed the effect of memory resources on the learning behavior of the incremental BayesCat model (Section 4.5.2). We examined the effect of the number of particles available to the particle filters, as well as the effect of rejuvenation.

Across experimental settings, we showed that the one-particle filter is outperformed by filters which explore multiple hypotheses simultaneously. Our results thus suggest that having access to one hypothesis at a time, during learning, is not sufficient for our category acquisition task. However, we also observe that an increased number of particles does not necessarily lead to increased performance. A filter with five particles is able to substantially outperform a filter with one particle, while not being much worse than a filter with 100 particles. In the literature it has been argued, following the *singularity principle*, that humans have a strong tendency to consider only the one most likely category in reasoning at any time (Evans, 2007; Murphy et al., 2012), which is at odds with our observations above. However, we point out that BayesCat is a model of child category *acquisition* whereas the research investigates *categorization* of objects in lab experiments with adult participants. It would be interesting investigate whether the singularity principle holds in a learning setting similar to ours.

We further showed that our model resembles human learning in the sense that the learner's uncertainty decreases over time, as measured by the frequency of resampling. Intuitively, would expect that early state representations in human learning are more uncertain than later ones. With more observed stimuli, the learnt knowledge should become more stable, and revisions of the knowledge state should occur less frequently. We observe this behavior in our particle filters as well: in the initial learning phase resampling is very frequent, but the frequency decreases over time (cf., Figure 4.10).

Our final set of experiments (Section 4.5.3) compared two models in their ability to rank concepts with respect to typicality, against a human created gold standard. We showed that our model successfully captured the typicality of concepts within a given category. The typicality ratings produced by BayesCat (Table 4.8) largely correspond to human intuitions. We should also point out that this is a large-scale study over hundreds of concepts. Previous work on the same task has only used a few dozens (Storms et al., 2000; Voorspoels et al., 2008; Connel and Ramscar, 2001). BayesCat outperforms a simpler vector space model which is nonetheless non-incremental. Our model learns statistical information about observed concepts incrementally, whereas the vector spaced model has all information available at once for constructing concept representations. BayesCat exhibits better typicality performance, which suggests that (a) the learnt concept representations are meaningful and (b) the incremental learning procedure does not put the model at disadvantage. Finally, we should note that BayesCat was not optimized or tuned for the typicality rating task in any way. Typicality follows naturally from the model structure without any additional assumptions on the task or learning strategy.

A common problem for models based on co-occurrence patterns in text (like BayesCat) is the fact that word type distributions follow a power law, and are consequently highly skewed. The induced information from raw text tends to be dominated by function words which occur with high frequency, but do not carry meaning themselves. While sophisticated priors can help alleviate the problem (Wallach et al., 2009), a more common strategy is to filter very high-frequent and low frequent words from the input to reduce the ‘skewness’ of the data. We apply this filtering to all input corpora used in experiments in this thesis. Without input filtering, we would expect the interpretability and relevance of the learnt categories and, in particular, features to decrease. Especially the incrementally updated representations induced by the particle filter would be likely dominated by high-frequency words early on. In addition, vocabulary filtering reduces the dimensionality of some of the model distributions, which ensures tractability of learning and inference. From a cognitive point of view, input filtering can be interpreted as an approximation of attention: through information beyond pure speech, such as prosody or gaze as well as cross-situational experience, children’s attention is guided to the relevant words and objects in their environment (Dominey and Dodane, 2004), i.e., meaning-bearing words in the linguistic input.

4.6 Summary

In this chapter we have presented BayesCat, a Bayesian model of category acquisition. Our model learns to group linguistic concepts into categories as well as their features (i.e., context words associated with them). Category learning is performed incrementally, using a particle filtering algorithm which is a natural choice for modeling sequential aspects of language learning. Our experiments were designed to answer several questions with respect to the robustness of the proposed model, the quality of its output, and adopted learning mechanism. (1) How do the induced categories fare against gold standard categories? (2) Are there performance differences between BayesCat and Chinese Whispers, given that the two models adopt distinct mechanisms for representing lexical meaning and learning semantic categories? (3) Does our learning mechanism predict human performance and is it cognitively plausible? We now summarize our findings in the light of the above questions.

Firstly, we observe that our incremental model learns plausible linguistic categories when compared against the gold standard. Secondly, these categories are qualitatively better when evaluated against Chinese Whispers, a closely related graph-based incremental algorithm. Thirdly, analysis of the model's output shows that it simulates category learning in two important ways, it consistently improves over time and can additionally acquire category features. Overall, our model has a more cognitively plausible learning mechanism compared to CW, and is more expressive, as it can simulate both category and feature learning. Although CW ultimately yields some meaningful categories, it does not acquire any knowledge pertaining to their features. This is somewhat unrealistic given that humans are good at inferring missing features for unknown categories (Anderson, 1991). It is also symptomatic of the nature of the algorithm which does not have an explicit learning mechanism. Each node in the graph iteratively adopts (in random order) the strongest class in its neighborhood (i.e., the set of nodes with which it shares an edge). We also explored how memory resources affect the learner's performance and showed that it is beneficial to entertain multiple hypotheses (i.e., numbers of particles) during learning. Furthermore, our model is able to revisit past decisions via rejuvenation. We experimentally showed that the learner revisits past decisions more frequently in the initial stages of learning when knowledge is being acquired and there is more uncertainty. Our final experiment showed that our model performs well on a typicality rating task when compared against a non-incremental semantic space.

In our experiments, the BayesCat model learnt with Gibbs sampling yielded a categorization which is a closer fit to the cognitive gold standard compared to the particle filter. Does this mean that the Gibbs sampler is a more plausible algorithm? From a learning perspective, the answer is no: aside from the fact that humans acquire knowledge incrementally, processing limitations do not permit revisiting past decisions exhaustively, by iterating over past experiences, as is the case for the Gibbs sampler. In view of this limitation, the incremental particle filters perform competitively throughout our experiments.

Overall, our results highlight the advantages of the Bayesian framework for modeling inductive problems and their learning mechanisms. Particle filters in particular suggest a class of psychologically plausible procedures for learning under cognitive constraints (e.g., memory or computational limitations). Although our experiments focused exclusively on categorization, we believe that some of the inference algorithms employed here could be easily adapted to other cognitive tasks such as word learning, word segmentation, phonetic learning, and lexical category acquisition. Importantly, we have shown that incremental learning in a Bayesian setting is robust and scalable in the face of large volumes of data, and the resulting models perform competitively compared to batch optimal learners.

Taken together our results further provide support for the important role of *distributional information* in categorization. We have demonstrated that co-occurrence information can be used to model how categories are learnt. Moreover, our typicality experiments indicate that the responses people provide in typicality experiments are to a certain extent reflective of the distributional properties of the linguistic environments in which concepts are found. Although our focus in this chapter has been primarily on the learning mechanisms of categorization, our experiments suggest that language itself is part of the environment that determines conceptual behavior. Furthermore, the fact that our models learn plausible categorizations from linguistic data alone would seem to indicate that information relating to the perceptual experience of objects and artifacts is encoded (albeit implicitly) in linguistic experience. In future work, it would be interesting to tease the contributions of linguistic and perceptual experience apart. It seems likely that no grounding is necessary for some concepts (or categories), whereas for others grounding is essential.

In the future we would also like to relax some of our simplifying assumptions regarding the learning environment which considers a single modality, namely language. It is possible to augment the set of features our model is exposed to with information from other modalities, such as the visual features of a scene, while leaving the model structure and learning algorithm unchanged. Another potential extension would involve augmenting the learning domain of the BayesCat model. In our experiments, the set of target concepts was constrained to those present in our gold standard. This was expedient for evaluation purposes, however there is no inherent limitation in the model which restricts its application to a specific domain or number of words. It would be interesting to see whether the features learned by a model trained on a larger set of target words differ qualitatively from those inferred from more limited domains.

We showed that BayesCat induces meaningful categories under a cognitively plausible learning mechanism. However, it learns *unstructured* bags-of-features for each category. This is in conflict with results from prior research which suggest that humans represent category knowledge in structured ways, resembling the structure of the world they represent (Murphy and Medin, 1985; McRae et al., 2005). The features learnt by BayesCat emerge as a by-product of category acquisition – they are not optimized themselves during learning. Experimental evidence suggests, however, that categories and features are learnt jointly, in a single process and mutually influence each other (Schyns and Rodet, 1997; Goldstone et al., 2001). We address these shortcomings in the following chapter, where we develop a Bayesian model which learns categories and structured featural representation jointly in a single process.

Chapter 5

Joint Acquisition of Categories and their Structured Feature Representations

Categorization is one of the most basic cognitive functions. It allows individuals to organize subjective experience of their environment by structuring its contents. This ability to group different concepts into the same category based on their common characteristics underlies major cognitive activities such as perception, learning, and the use of language. Global semantic categories (such as FURNITURE or ANIMAL) are shared among members of societies, and influence how we perceive, interact with, and argue about the world.

Given its fundamental importance, categorization is one of the most studied problems in cognitive science. The literature is rife with theoretical and experimental accounts, as well as modeling simulations focusing on the emergence, representation, and learning of categories. Most theories assume that basic level concepts such as *dog* or *chair* are characterized by features such as {barks, used-for-sitting}, and are grouped into categories based on those features. Although the precise grouping mechanism has been subject to considerable debate (including arguments in favor of *exemplars* (Nosofsky, 1988), *prototypes* (Rosch, 1973), and category *utility* (Corter and Gluck, 1992)), it is fairly uncontroversial that categories are associated with featural representations.

Less effort has been dedicated to the question of where those features come from. Much theoretical and computational work on categorization assumes a fixed, readily available set of adequate features for categorization-related tasks. Recent theoretical work, however, has challenged these assumptions. Experimental studies show that the development of categories and feature learning mutually influence each other (Goldstone et al., 2001; Schyns and Rodet, 1997; Spalding and Ross, 2000): concepts are categorized based on their features, but the perception of features is influenced by already established categories, and, like categories, features evolve over time. There is also evidence that features such as {barks, runs} are grouped into types like `behavior` (Ahn, 1998; McRae et al., 2005; Wu and Barsalou, 2009), and the distribution of feature types varies across categories (McRae and Cree, 2002). For instance, living things such as ANIMALS have characteristic `behavior`, whereas artifacts such as TOOLS have characteristic `functions`, and both categories have characteristic `appearance`.

Previously proposed models for category learning have largely considered the problems of category and feature learning in isolation, focusing either on category learning given a limited set of simplistic features (Anderson, 1991; Sanborn et al., 2006) or feature learning (Austerweil and Griffiths, 2013; Baroni et al., 2010; Kelly et al., 2014), but not both; or they learn from restricted, task-specific data sets (Shafto et al., 2011). Moreover, influential models of categorization (such as Anderson (1991)'s rational model of categorization, or ALCOVE (Kruschke, 1992) among many others) rely on the availability of a pre-defined and fixed set of informative features associated with every observed concept, such that "[...] the modeler, not the model, [... chooses] the appropriate features for the considered categorizations" (Schyns and Rodet, 1997, p. 684). Such models furthermore assume that features are independent and combine linearly rather than being correlated or structured.

Our own BayesCat model, introduced in Chapter 4, learns categories from more realistic input data in the sense that it is exposed to occurrences of concept mentions in their natural language context, which may be noisy or contain information irrelevant to the concept. BayesCat *learns* relevant features as sets of terms which are highly associated with specific categories from unfiltered input. Nevertheless, these features are (a) flat, unstructured sets; and (b) emerge as a by-product of the category learning process, but are not optimized themselves. In this chapter, we address these shortcomings and tackle the problem of *jointly* learning categories and their *structured* representations.

We induce categories (e.g., ANIMALS) and their feature types (e.g., `behavior`) from observations of target concepts (e.g., *lion*, *dog*, *cow*) and their co-occurring contexts (e.g., {eats, sleeps, large}). Specifically, our model induces a set of categories and their representations in a single process by learning (a) categories as clusters of concepts; (b) feature types as probability distributions over context words; and (c) category-feature type associations as category-specific distributions over feature types.

We apply our model to large-scale collections of encyclopedic text, assuming that featural information is particularly explicit in this data set. Evaluation results show that our cognitively motivated joint model learns accurate categories and feature types, achieving results competitive with highly engineered approaches focusing exclusively on feature learning. In line with our BayesCat evaluation in Chapter 4, we also investigate the behavior of our model under cognitively more plausible learning conditions. We (a) expose our model to data resembling the input a child has access to when learning categories and their features; and (b) model the human learning process more faithfully through an incremental learning algorithm. Our model observes training data sequentially and is subject to cognitive constraints in terms of memory limitations. We show that our model acquires meaningful categories and features from child-directed language, and analyze the influence of memory constraints on the learning process.

We begin this Chapter with a review of previous work (Section 5.1). We continue with a detailed description of our joint model (Section 5.2) and introduce two learning algorithms: an "ideal" batch learner, and an incremental learner (particle filter). Section 5.3 presents our large-scale evaluation based on encyclopedic data. In Section 5.4, we evaluate our incremental learner on child-directed language. We discuss our findings in Section 5.5 and draw conclusions from our results.

5.1 Categories and Structured Featural Representations

This section provides an overview of prior research in cognitive science which challenges the assumptions underlying many theories and models of categorization. We review work supporting the claims that (a) featural representations are structured into types of features and that the distribution of those types is category-specific (Section 5.1.1); and (b) that category and feature acquisition are a joint process and the two parts exert a mutual influence (Section 5.1.2). We also review relevant computational models of human category and feature learning from cognitive data and systems for feature extraction from text (Section 5.1.3).

5.1.1 Featural Representations of Concepts and Categories

Even though much empirical research glosses over this observation, there is strong evidence that human conceptual representations are structured (see Rips et al. (2012) for a recent critique and overview of cognitive studies of categorization). Categories mentally represent the complex structure of the environment. They allow to make inferences about concepts or categories that go beyond their perceived similarities capturing abstract and potentially complex properties of categories (for example the `nutritional value` of FOOD items, or the `emotional benefits` of PETS). Much research on human categorization is based on laboratory experiments where subjects are presented with artificial stimuli represented by a restricted set of task-relevant features. Observations of *natural* concepts, however, are often noisy or incomplete so that a notion of systematic relations among features might be more beneficial here than under artificial conditions in the lab (Malt and Smith, 1984). The authors also discuss that structure among features might be particularly important for superordinate level categories such as ANIMAL since they are represented by a much more heterogeneous set of features than basic level categories (e.g., *dog* or *fish*). Our model induces structured feature representations for superordinate level categories.

The existence of structured features has received support through behavioral results from a variety of categorization related tasks, such as typicality rating (Malt and Smith, 1984) or category-based inductive inference (Heit and Rubinstein, 1994; Spalding and Ross, 2000). Experimental evidence suggests that categories which on the surface do not seem to contain a coherent set of members (e.g., the category PETS) are represented by an underlying set of abstract features which explain the coherence of the category (e.g., {keeps_company, lives_in_the_house}). Subjects' categorical inferences suggest that observed surface features of category members systematically activate associated underlying features of newly observed concepts of the same category (Spalding and Ross, 2000).

Varying the types of available features (e.g., providing `functional` information in addition to objects' `appearance`) leads to different categorization behavior both in adults (Heit and Rubinstein, 1994) and children (Trauble and Pauen, 2007), and different feature types vary in their predictive value across categories. Jones et al. (1991) show that children as young as 2-3 years old possess a notion of relations between feature types and categories: otherwise identical stimuli were presented either with eyes (suggesting an ANIMATE category) or without. Children categorized stimuli with eyes based on both their `shape` and `texture`, and stimuli without eyes based on their `shape` only. Similarly, Macario (1991) showed that 2-4-year old children categorize FOOD items based on their `color`, however, TOYS are classified based on their `shape`.

Gelman and Markman (1986) and Gelman and O'Reilly (1988) showed that children are able to make category-related inferences that go beyond surface similarities of the involved concepts. For example, children were presented with two stimuli, e.g., (1) a tropical fish and (2) a large, gray dolphin. They were then presented a third stimulus, e.g., a large gray shark, which was perceptually similar to (2) but was labeled with the same name as (1) (i.e., fish) by the experimenter. Children made stronger inferences based on shared category labels than on surface similarity. Apart from investigating children's behavior in a category-related inference task, Gelman and O'Reilly (1988) also asked children to justify their inference decisions in an open-ended interview paradigm. Children were indeed aware of the fact that members of the same category shared important properties such as `behaviors` and `structures` irrespective of shared visual features.

Another line of work has investigated children's understanding of the fundamental *defining* characteristics that determine the category membership of concepts. Keil (1989) exposed participants to highly modified instances of ARTIFICIAL (e.g., *chair*) and NATURAL (e.g., *squirrel*) categories and asked whether the modified concept still belonged to the same category. He showed that surface manipulation changed category judgments of ARTIFACTS (glueing leg extensions onto a *chair* and sawing off its back changes its category into a *stool*), whereas similar manipulations do not change category membership judgments for animates (dying a *raccoon*'s fur, fluffing up its tail, and enabling it to release a smelly secretion when scared, leaves the animal's identity unchanged).[1] Conversely, modifying molecular structure changed category membership judgments for animates but not for artifacts (e.g., a *chair* made out of a material which is used for making *windsurfers* remains a *chair*; but a discovery of a fundamental mistake in the analysis of a plant's DNA will presumably change its category).

The structured nature of category features manifests itself in feature norms. Feature norms are verbalized lists of properties that humans associate with a particular con-

[1] These patterns reliably emerge from judgments collected from adults and older children, however, responses from younger children are less reliable and suggest the development of featural representations during concept and category acquisition in young children. We return to this phenomenon in detail in Chapter 6.

cept (McRae et al., 2005). Features collected in norming studies naturally fall into different types such as `behavior`, `appearance` or `function`. This suggests that structure also emerges from verbalized representations of concepts and features such as mentions in natural language corpora, as used as stimuli throughout this thesis. McRae et al. (2005) collected a large set of feature norm for more than 500 concepts in a multi-year study, and classified their feature norms using a variety of theoretically motivated schemata, including the feature type classification scheme developed in Wu and Barsalou (2009) and Barsalou (1999). Wu and Barsalou's work suggests that humans perform a "mental simulation" when describing a concept, they scan the mental image they create as well as situations associated with that image, and then verbalize it when producing features.

These rather theoretical motivations have received support from behavioral data collected from patients with brain damage and impaired categorization abilities (Warrington and Shallice, 1984; Humphreys and Forde, 2001). In their seminal study, Warrington and Shallice (1984) systematically observed category-specific impairments: groups of patients showed impaired categorization skills for ARTIFACTS, but also BODYPARTS, while performing normally on the categorization of ANIMATES and MUSICAL INSTRUMENTS, and vice versa. While these observations are difficult to explain on the category level, the authors suggest an interpretation on the featural level: ARTIFACTS and BODYPARTS, are strongly associated with `functional` features (such as, {used_for_eating, used_for_walking}). On the other hand ANIMATES, like MUSICAL INSTRUMENTS, are associated with `perceptual` or `behavioral` features (such as, {makes_sounds}). A local representation of the respective feature types in separate brain areas would provide an explanation of the observed behavior. Although the authors offer no empirical evidence for their category-feature association hypothesis, subsequent work revealed that descriptions of ARTIFACTS in dictionary entries contain more functional features, whereas description of ANIMATES contain more perceptual features (Farah and McClelland, 1991).

Further empirical evidence was provided by McRae and Cree (2002), who show that feature type information extracted from a large collection of feature norms (discussed above) explains not only the binary `perceptual` vs `functional` dichotomy found by Warrington and Shallice (1984), but also a number of additional category-specific deficits observed in brain-damage patients. Categories were represented in terms of feature types created from a large set of human-produced feature norms using Wu and

Barsalou (2009)'s coding scheme, which was developed and motivated independently of explaining categorization deficits. Categories were then clustered using average-link agglomerative clustering and groups of categories which were treated identically (either impaired or non-impaired) by patients emerged. The results support the centrality of feature types in category representations, providing further evidence that they underly category representations in the brain.

The model we present in this chapter aims to capture the evidence summarized above, and represent categories as *structured* sets of associated features. Category-specific features are structured into types which relate to a particular kind of property of a category (e.g., the `behavior` of ANIMALS). We also capture the observation that features are defining for different categories to a varying degree (Keil, 1989; McRae and Cree, 2002) in terms of category-feature type associations (e.g., the feature `function` is highly defining for (or associated with) the category ARTIFACT not for the category ANIMAL).

5.1.2 Joint Learning of Categories and their Features

Although the majority of models of categorization assume a fixed set of features underlying the category acquisition and categorization process, there is increasing evidence that "[...] a significant part of learning a category involves learning the features entering its representations." (Schyns and Rodet, 1997, p. 681). Experimental evidence suggests that not only do features underly the categorization process but features themselves are susceptible to change over time and can be modified by the categories which emerge. Evidence ranges from changing featural perception as a result of expert education (e.g., wine tasters or doctors learning to interpret X-ray images) to neurological evidence revealing enhanced neural activity in experts when presented with pictures of their area of expertise (see Goldstone et al. (2008) for an overview).

The influence of category learning on the perception and use of features has been studied extensively using visual stimuli of varying degree of naturalness and familiarity. Pevtzow and Goldstone (1994) experiment with drawings of 2-dimensional line segments, and show that participants who were exposed to categorization training prior to a feature identification task identified the presence of category-defining features faster than participants without prior training. Goldstone et al. (2001) and Goldstone and Steyvers (2001) presented participants with photographs of human faces, which were systematically manipulated. Participants were then asked to categorize the faces, and after the categorization task showed higher sensitivity to the features which were relevant for the categorization.

In contrast to the work discussed above which uses familiar concepts, Schyns and Rodet (1997); Schyns et al. (1998) computer-generate visually complex 2-d images of “Martian cells”, demanding substantial familiarization from their participants thus minimizing the influence of prior knowledge which enables them to study the emergence of features and change of their perception in isolation. Their experiments show that (a) knowing the categorization of perceptual stimuli changes the perceptual units on which the analysis of those stimuli are based; and (b) a change in the order of category learning (based on change of order of stimulus presentation) influences the perception of stimuli and leads to the emergence of different features.

Other work has explicitly investigated the *incremental* learning process of categories and the ways how evidence encountered later in the learning process changes category structure (Ross, 1997, 2000). In contrast to the work mentioned above, such experiments involve conceptual rather than perceptual stimuli: participants learn classes of diseases based on symptoms. They show that *using* learnt categories in categorization tasks alters category representations, both when learning and usage are interleaved but also when the learning process precedes usage.

Diaz and Ross (2006) investigate the incremental process of learning category structure. They show that interleaving inference questions on feature correlations with categorization tasks mutually improves performance on both tasks over time. While the previous experiment was conducted with adult participants, Bornstein and Mash (2010) show that 5-months old infants are capable of learning categories of unfamiliar objects on-line as new information becomes available.

Our model is the first model which learns categories and their features *jointly* in one process from naturalistic input data. We show that meaningful categories as well as relevant structured features emerge in an incremental manner, capturing characteristics of the human learning process.

5.1.3 Computational Models of Category and Feature Induction

We conclude our background section with a review of prior work on computational models of category and structured feature induction. We begin with an overview of cognitive models which aim to replicate behavioral data. Afterwards we review knowledge-heavier approaches for feature extraction from text corpora.

The problems of category formation and feature learning have been considered largely independently in the literature. Bayesian categorization models pioneered by Anderson (1991) and recently re-formalized by Sanborn et al. (2006) are aimed at replicating human behavior in small scale category acquisition studies, where a fixed set of simple (e.g., binary) features is assumed. Our BayesCat model presented in Chapter 4 of this thesis is similar in spirit, but was applied to large-scale corpora, while investigating incremental learning in the context of child category acquisition (see also Fountain and Lapata (2011) for a non-Bayesian approach). BayesCat associates sets of features with categories as a by-product of the learning process, however these feature sets are independent across categories and are not optimized during learning.

A variety of cognitively motivated Bayesian models have been proposed for the acquisition of complex domain knowledge. Shafto et al. (2011) present a joint model of category and feature acquisition in the context of cross-categorization, i.e., the phenomenon that concepts are simultaneously organized into several categorizations and the particular category (and features) that are relevant depend on the context (e.g., concepts of the category FOOD can be organized based on their `nutritional` or `perceptual` properties). They develop Bayesian models for category and feature learning and find that only a joint model for both processes explains the emergence of cross-cutting categories. In addition to a series of experiments based on small data sets (comprising 8 concepts, 6 features and 2 categorization systems), Shafto et al. (2011) also evaluate their model in a more naturalistic setting on a cross-categorization task of ANIMALS and MUSICAL INSTRUMENTS based on human-produced feature norms. This work is similar to our work in terms of the attempt to learn categories and features jointly. However, while Shafto et al. (2011) present their model with category-specific data sets tailored towards their learning objective, we are interested in *acquiring* categories and structured associated features jointly from thematically unconstrained corpora of natural text.

Another line of work (Kemp et al., 2003; Perfors et al., 2005) models the joint learning of relevant features and domain-specific feature type biases in children. They focus on the acquisition of domain-specific representational structures (such as hierarchies or clusters) and discuss results in the context of word learning. In contrast to our work, their model assumes a priori established categories (such as FOOD and ANIMALS), and learns from task-specific data representations in the form of objects described by a limited set of relevant features (however, a weighting of those features is learnt). Perfors and Tenenbaum (2009) present a Bayesian model which simultaneously learns categories (i.e., groupings of concepts based on shared features) and *learns to learn* categories (i.e., abstract knowledge about kinds of featural regularities that characterize a category). They compare their model predictions against behavioral data from adult participants, which limits the scope of their experiments to small data sets e.g., of artificial stimuli with a restricted number of abstract features. In addition to the differences in the training data, the models discussed so far were not tested under cognitively motivated learning conditions, e.g., by using incremental learning algorithms.

The ability to automatically extract feature-like information for concepts from text would facilitate the laborious process of feature norming and improve the coverage concepts and their features. A few approaches to feature learning from textual corpora exist, and they have primarily focused on emulating or complementing norming studies by automatically extracting norm-like properties from corpora (e.g., *elephant* `has-trunk`, *scissors* `used-for-cutting`). Steyvers (2010) use a flavor of topic models to augment data sets of human-produced feature norms. While vanilla topic models (Blei et al., 2003) represent documents as sets of corpus-induced topics, Steyvers (2010) additionally use topics derived from the feature norms. The learnt topics yield useful extensions of the original feature norms, with properties that were previously not covered, suggesting that corpora are an appropriate resource for augmenting feature norms of concepts.

Another line of research concerns entirely text-based feature extraction. A common theme in this line of work is the use of pre-defined syntactic patterns (Baroni et al., 2010), or manually created rules specifying possible connection paths of concepts to features in dependency trees (Devereux et al., 2009; Kelly et al., 2014). While the set of syntactic patterns pre-defines the relation types the system can capture, the latter approach can extract features which are a priori unlimited in their relation to the target concept. Once extracted, the features are typically weighted using statistical measures of association in order to filter out noisy instances. Similar to our own work, the motivation underlying these models is large-scale unsupervised feature extraction from text. These systems are not cognitively motivated *acquisition* models, however, due to (a) the assumption of involved prior knowledge (such as syntactic parses or manually defined patterns), and (b) the two stage extraction-and-filtering process which they adopt. Humans arguably do not first learn a large set of potential features for concepts, before they infer their relevance. The systems discussed above learn features for individual concepts rather than categories.

To our knowledge, we propose the first Bayesian model that jointly learns categories and their features from naturalistic large-scale input data. Our model is knowledge-lean, it learns from raw text in a single process, without relying on parsing resources, manually crafted rule patterns, or post-processing steps. Our work also differs from approaches which combine topic models with human-produced feature norms (Steyvers, 2010). Our aim is not to boost the generalization performance of a topic model, rather we investigate how both categories and features can be jointly learnt from data.

5.2 A Bayesian Model for Joint Learning of Categories and their Features

This section presents our Bayesian model of category and feature induction (henceforth, BCF). We give an intuitive overview of BCF, before we formally derive our model. Afterwards, we derive two approximate learning algorithms: a Gibbs sampler for batch learning (Section 5.2.1) and an incremental particle filter (Section 5.2.2).

Intuition BCF is a joint model for learning categories and their featural representation. In a single process, BCF acquires (a) categories (e.g., {ANIMAL, VEHICLE, TOOL}) of concepts (e.g., {*cat, car, drill*}), (b) feature types (e.g., `appearance`, `diet`, `utility`), and (c) associations between categories and feature types. Specifically, it infers one global set of feature types which is shared across categories (e.g., ANIMALS and VEHICLES can be described in terms of `colors`). However, categories differ in their strength of association with feature types (e.g., the feature type `function` will be highly associated with TOOLS, but less so with ANIMALS). BCF jointly optimizes categories and their featural representation: the learning objective is to obtain a set of meaningful categories, each characterized by relevant and thematically coher-

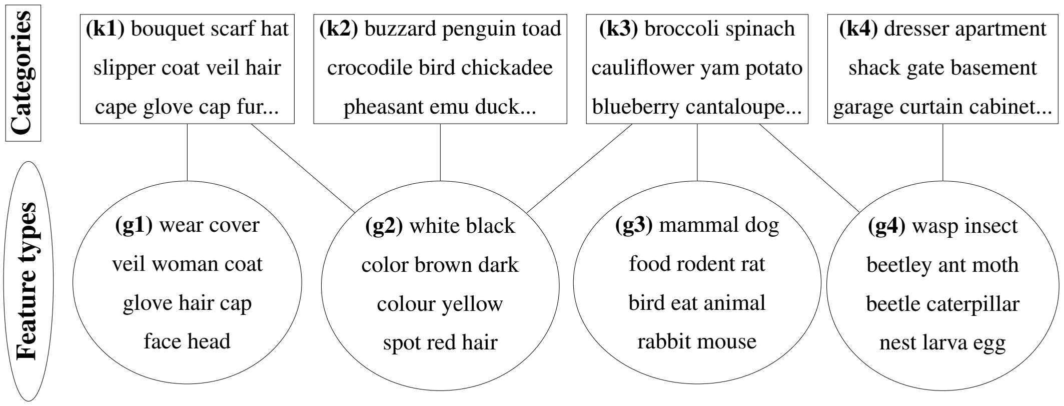

Categories
(k1) bouquet scarf hat slipper coat veil hair cape glove cap fur...
(k2) buzzard penguin toad crocodile bird chickadee pheasant emu duck...
(k3) broccoli spinach cauliflower yam potato blueberry cantaloupe...
(k4) dresser apartment shack gate basement garage curtain cabinet...
Feature types
(g1) wear cover veil woman coat glove hair cap face head
(g2) white black color brown dark colour yellow spot red hair
(g3) mammal dog food rodent rat bird eat animal rabbit mouse
(g4) wasp insect beetley ant moth beetle caterpillar nest larva egg

Figure 5.1: Examples of categories (top) and feature types (bottom) inferred by the BCF model from a corpus of encyclopedic text. Connecting lines indicate a strong association between the category and the respective feature type.

ent feature types. Input to our model is a collection of natural language text stimuli, each of which consists of a mention of target concept, within its local linguistic context. We treat each stimulus as an observation of the concept: the word referring to the concept as an instance of the concept itself, and its context words as a representation of its features. The set of target concepts is fixed, however, the set of context words (i.e., features) is potentially unbounded and determined by our input text corpora.

We assume that each concept belongs to a single category. We further assume that each input stimulus refers to exactly one underlying feature type. Our goal in inference is to assign a feature type to each input, as well as a category to each concept type. Specifically, the occurrences of a concept will be assigned a category based on how similar the concept's associated feature types are compared to the feature types associated with any potential category. Simultaneously, upon observing a stimulus (i.e., a concept in context), the model assigns the context to a particular feature type based on its probability under all potential feature types, and the prior probability of observing that feature type with the stimulus concept's assigned category. From a cognitive point-of-view this is intuitive: relevant features are triggered by category membership: *cats* and *dogs* might be discussed in terms of their `diet`, but *chairs* and *tables* are not; their `material` may however be a relevant featural aspect. Conversely, if we encounter an instance of a previously unknown concept which is described in terms of its `diet`, we may infer that it belongs to the category ANIMAL with a higher probability than to the category FURNITURE.

Figure 5.1 illustrates example output produced by our model, in terms of learnt cat-

symbol	explanation	
$d \in \{1..D\}$	stimulus	(e.g., “This dog likes to catch balls.”)
$c \in \{1..C\}$	concept mention in a stimulus	
$i \in \{1..I\}$	context word positions in a stimulus	
$\ell \in \{1..\mathcal{L}\}$	concept types	(e.g., *cat, dog, chair, table*)
$f \in \{1..V\}$	features	(e.g., {runs, barks, eats} {red, made_of_wood})
$k \in \{1..K\}$	categories	(e.g., ANIMAL, FURNITURE)
$g \in \{1..G\}$	feature types	(e.g., `behavior, appearance`)
θ	K-dimensional parameter vector of category distribution	
$\{\mu_k\}_{k=1}^{K}$	G-dimensional parameter vectors of feature type distributions	
$\{\phi_g\}_{g=1}^{G}$	V-dimensional parameter vectors of word distributions	

Table 5.1: Notational overview of the BCF model (the category and feature type labels are provided for illustration; BCF is an unsupervised clustering model which induces unlabeled categories and feature types).

egories, learnt feature types and their mutual associations. Connecting lines indicate category-feature type associations. Feature types are shared across categories, for example the categories CLOTHING (k1), BIRDS (k2), and FOOD (k3) are all associated with feature type `color` (g2).

Model Description We now describe the BCF model more formally. Our model is parameterized with respect to the number of categories K it can infer, as well as the number of global feature types G that are available across categories. A distribution over feature types is inferred for each category. We furthermore specify the set of $\mathcal{L}$ target concepts a priori, and provide an input corpus which consists of stimuli covering all and only these target concepts.

A notational overview is provided in Table 5.1 The generative story of our model is displayed in Figure 5.2a, and Figure 5.2b shows the plate diagram representation of BCF. The generative story proceeds as follows. We assume a global multinomial distribution over categories $Mult(\theta)$. Its parameter vector θ is drawn from a symmetric Dirichlet distribution with hyperparameter α. For each concept type $\ell = [1...\mathcal{L}]$, we draw a category k^{ℓ} from $Mult(\theta)$. For each category k, we assume an independent set of multinomial parameters over feature types μ_k, drawn from a symmetric Dirichlet distribution with hyperparameter β. Finally, for each feature type g, we draw a

(a) Generative story of BCF.

```
Generate category distribution, $\theta \sim Dir(\alpha)$

for concept type $\ell = 1..\mathcal{L}$ do
    Generate category, $k^{\ell} \sim Mult(\theta)$

for category $k = 1..K$ do
    Generate feature type distribution, $\mu_k \sim Dir(\beta)$

for feature type $g = 1..G$ do
    Generate feature distribution, $\phi_g \sim Dir(\gamma)$

for stimulus $d = 1..D$ do
    Observe concept $c^d$ and retrieve category $k^{c^d}$
    Generate a feature type, $g^d \sim Mult(\mu_{k^{c^d}})$
    for feature position $i = 1..I$ do
        Generate a feature $f_{d,i} \sim Mult(\phi_{g^d})$
```

(b) Plate diagram of BCF.

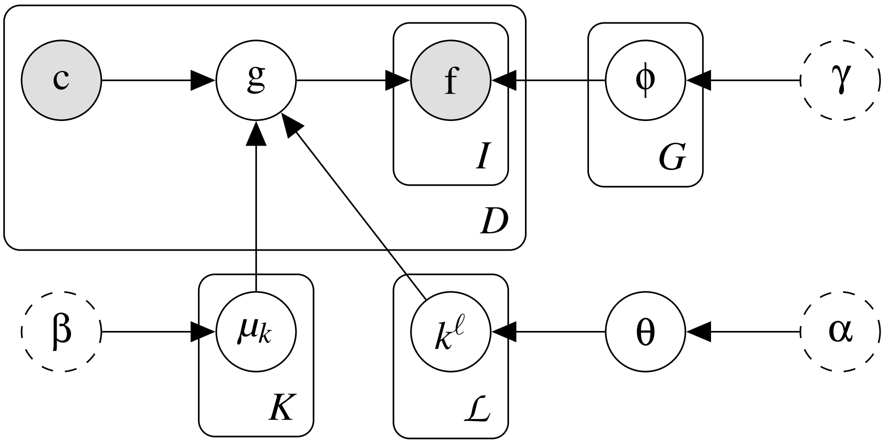


Figure 5.2: Top (a): The generative story of the BCF model. Observations f and latent labels k and g are drawn from Multinomial distributions ($Mult$). Parameters for the multinomial distributions are drawn from Dirichlet distributions (Dir). Bottom (b): The plate diagram of the BCF model. Shaded nodes indicate observed variables, clear nodes denote latent variables, and dotted nodes indicate constant hyperparameters.

multinomial distribution over features $Mult(\phi_g)$ from a symmetric Dirichlet distribution with hyperparameter γ. With these global assignments in place, we can generate sets of stimuli $d = [1...D]$ as follows: we first retrieve the category k^{c^d} of an observed concept c^d; we then generate a feature type g^d from the category's feature type distribution $Mult(\mu_{k^{c^d}})$; and finally, for each position $i = [1...I]$ we generate feature $f_{d,i}$ from the feature type-specific feature distribution $Mult(\phi_{g^d})$.

The joint probability of the model over latent categories, latent feature types, model parameters, and data factorizes as:

$$\begin{aligned} &p(g,f,\mu,\phi,\theta,k|c,\alpha,\beta,\gamma) = \\ &p(\theta|\alpha)\prod_{\ell} p(k^{\ell}|\theta)\prod_{k} p(\mu_k|\beta)\prod_{g} p(\phi_g|\gamma)\prod_{d} p(g^d|\mu_{k^{c^d}})\prod_{i} p(f^{d,i}|\phi_{g^d}). \end{aligned} \tag{5.1}$$

Since we use conjugate priors throughout, we can integrate out the model parameters analytically, and perform inference only over the latent variables, namely the category and feature type labels associated with the stimuli (see Sections 3.2.2.1 and 3.2.2.2 for a mathematical discussion of this approach).

To sum up, our model takes as input a text corpus of concept mentions in local context, and infers a concept categorization, a global set of feature types, as well as a distribution over feature types per category. After integrating out model parameters where possible, we infer two sets of latent variables:

(1) feature type-assignments to each stimulus $\{g\}^D$,

(2) category-assignments to each concept type $\{k\}^{\mathcal{L}}$.

The next two sections introduce a batch learning algorithm (a Gibbs sampler; Section 5.2.1) as well as a cognitively motivated incremental learning algorithm (a particle filter; Section 5.2.2) for approximate estimation of these parameters.

5.2.1 Batch Learning

Exact inference in the BCF model is intractable, so we turn to approximate posterior inference to discover the distribution over value assignments to latent variables given the observed data. In this section we introduce a Gibbs sampling algorithm (Geman and Geman, 1984) which is a Markov chain Monte Carlo algorithm which iteratively re-assigns single variables based on the current assignments of all other variables. We

Algorithm 4 The Gibbs sampling algorithm for the BCF model.

```
Input: model with randomly initialized parameters.
Output: posterior estimate of $\theta, \phi$, and $\mu$.
repeat
    for stimulus $d$ do                    ▷ Sample stimulus-feature type assignments
        decrement stimulus $d$-related counts
            $g^d \sim p(g^d_{k^{c^d}} = i | \mathbf{g}^{-d}_{k^{c^d}}, \mathbf{f}^-, k^{c^d}, \beta, \gamma)$        Equation (5.3)
        update stimulus $d$-related counts
    for concept $c$ do                     ▷ Sample concept-category assignments
        retrieve category $k^c$
        decrement concept $c$-related counts
            $k^c \sim p(k^\ell = j | \mathbf{g}_{k^\ell}, \mathbf{k}^-, \alpha, \beta)$        Equation (5.5)
        update concept $c$-related counts
until convergence
```

discussed the theory underlying Markov chain Monte Carlo and the Gibbs sampler in detail in Section 3.3.2 and describe here the particular instantiation of the algorithm for our BCF model. The sampling procedure is summarized in Algorithm 4. The Gibbs sampler repeatedly iterates over the training data set and resamples values of the latent variables. One Gibbs iteration for our model consists of two blocks:

Resampling stimulus-feature type assignments. In the first block we iterate over the input stimuli d, and resample each stimulus-feature type assignment g^d from its full conditional posterior distribution over feature types conditioned on (a) the values assigned to all other latent variables unrelated to the current variable of interest, i.e, all features except those in stimulus d, $\left(\mathbf{f}^-\right)$, and all stimulus-feature type assignments except the one to stimulus d, $\left(\mathbf{g}^{-d}_{k^{c^d}}\right)$; (b) the category currently assigned to d's target concept c, $\left(k^{c^d}\right)$; and (c) the relevant hyperparameters $\left(\beta, \gamma\right)$:

$$p(g^d_{k^{c^d}} = i | \mathbf{g}^{-d}_{k^{c^d}}, \mathbf{f}^-,\ k^{c^d} = j,\ \beta,\ \gamma) \tag{5.2}$$

$$= p(g^d_{k^{c^d}} = i | \mathbf{g}^{-d}_{k^{c^d}}, k^{c^d} = j, \beta) \quad \times \quad p(f^d | \mathbf{f}^-, g^d_{k^{c^d}} = i, \gamma) \tag{5.3}$$

$$\propto \frac{(n^j_i + \beta)}{(\sum_i n^j_i + \beta)} \quad \times \quad \frac{\prod_v \prod_{a=1}^{f_v} (n^i_v + \gamma + a)}{\prod_{a=1}^{f_*} (\sum_v n^i_v + \gamma + a)}. \tag{5.4}$$

The factorization of the posterior distribution in (5.3) follows from the model structure as described above and shown in the plate diagram in Figure 5.2b. The posterior distribution factors into the probability of a particular feature type i and the probability of the observed features in the stimulus given that feature type. Because of the Dirichlet-Multinomial conjugacy in our model, these two distributions can be straightforwardly computed using only the counts of current value-assignments to all variables in the model except the ones currently resampled (equation (5.4)):[2] the probability of a hypothetical feature type i is proportional to the number of times it has assigned previously to stimuli with observed category j, n^j_i, smoothed by the Dirichlet parameter β. Similarly, the probability of the observed features of stimulus d under hypothetical feature type i is proportional to the number of times each individual feature v in d has been observed under feature type i, n^i_v (smoothed by the Dirichlet parameter γ). In the second term in (5.4), f_v refers to the count of any particular feature v in stimulus d, and f_* refers to the number of features in d (irrespective of their value).

We compute the (unnormalized) probabilities of individual hypothetical feature types i as explained above. These values are then normalized and a new feature type is sampled from the resulting distribution.

Resampling concept-category assignments. The second block of our Gibbs sampler performs a sweep over all concept types $\ell \in \{1...\mathcal{L}\}$, and resamples each concept type ℓ's category assignment k^ℓ. Similarly to the process described above, the new category assignment of concept ℓ is resampled from its full conditional distribution over categories conditioned on (a) all concept-category assignments except for k^ℓ, $\left(\mathbf{k}^-\right)$; (b) the feature type assignments relevant to concept ℓ, $\left(\mathbf{g}^-_{k^\ell}\right)$; and (c) all relevant hyperparameters $\left(\alpha, \beta\right)$:

$$\begin{aligned} p(k^\ell = j|\mathbf{g}^-_{k^\ell},\, \mathbf{k}^-,\, \alpha,\, \beta) &= p(k^\ell = j|\mathbf{k}^-, \alpha) & \times \quad & p(\mathbf{g}_{k^\ell}|\mathbf{g}^-_{k^\ell}, k^\ell = j, \beta), \qquad (5.5) \\ &\propto (n^j + \alpha) & \times \quad & \frac{\prod_g \prod_{a=1}^{f^\ell_g} (n^j_g + \beta + a)}{\prod_{a=1}^{f^\ell_*} (\sum_g n^j_g + \beta + a)}. \qquad (5.6) \end{aligned}$$

Based on the independence assumptions in our model, this probability factorizes into the prior probability of hypothetical category j and the probability of feature types observed with concept ℓ under the hypothetical category j (equation (5.5)). Like above

[2] Please refer to Section 3.3.2.3 (p. 47) for a mathematical explanation of this result, and to Appendix A for a detailed derivation.

these probabilities can be computed purely based on counts of variable-assignments in the current sampler state (equation (5.6)). In the second term of (5.6), f_g^ℓ refers to the number of times feature type g was assigned to a stimulus containing concept type ℓ, and f_*^ℓ to the number of stimuli containing ℓ (irrespective of the assigned feature type).

Using the procedure described above we compute an (unnormalized) probability for each hypothetical category, normalize the probabilities and resample concept ℓ's category k^ℓ from the resulting distribution.

5.2.2 Incremental Learning

The Gibbs sampler introduced above stores the complete training data set, and passes over it repeatedly, approximating the target posterior distribution in an iterative fashion. This process does not resemble the nature of human learning. Humans learn incrementally, updating knowledge on-the-fly with information observed in the environment (Diaz and Ross, 2006), and are subject to memory limitations so that information observed in the environment is not stored for future processing in its entirety (Levy et al., 2009).

Here, we derive a particle filter (Doucet et al., 2001), an incremental learning algorithm, which instills in our BCF model a cognitively plausible learning mechanism: the training data are presented in an incremental fashion, stimulus by stimulus, and re-vision of previously encountered data is only possible to a limited extent. As discussed previously in this thesis (Section 3.3.3, Section 4.3.2), particle filters are an instantiation of the sequential Monte Carlo estimation framework which maintain an approximation of the target distribution through a set of hypotheses, and update these samples incrementally in real-time as novel information becomes available. Particle filters allow to flexibly adapt the memory capacity of the algorithm by varying the number of samples N maintained from the posterior distribution (particles). With an increasing number of particles, the filter can represent the probability space increasingly accurately (with the number of particles approaching infinity, the approximation is guaranteed to converge towards the target posterior distribution).

The process particle filtering for the BCF model is schematically illustrated in Figure 5.3 (a). Each particle maintains a sample as concrete instantiation of a categorization and featural representation. Figure 5.3 (b) illustrates the information contained in

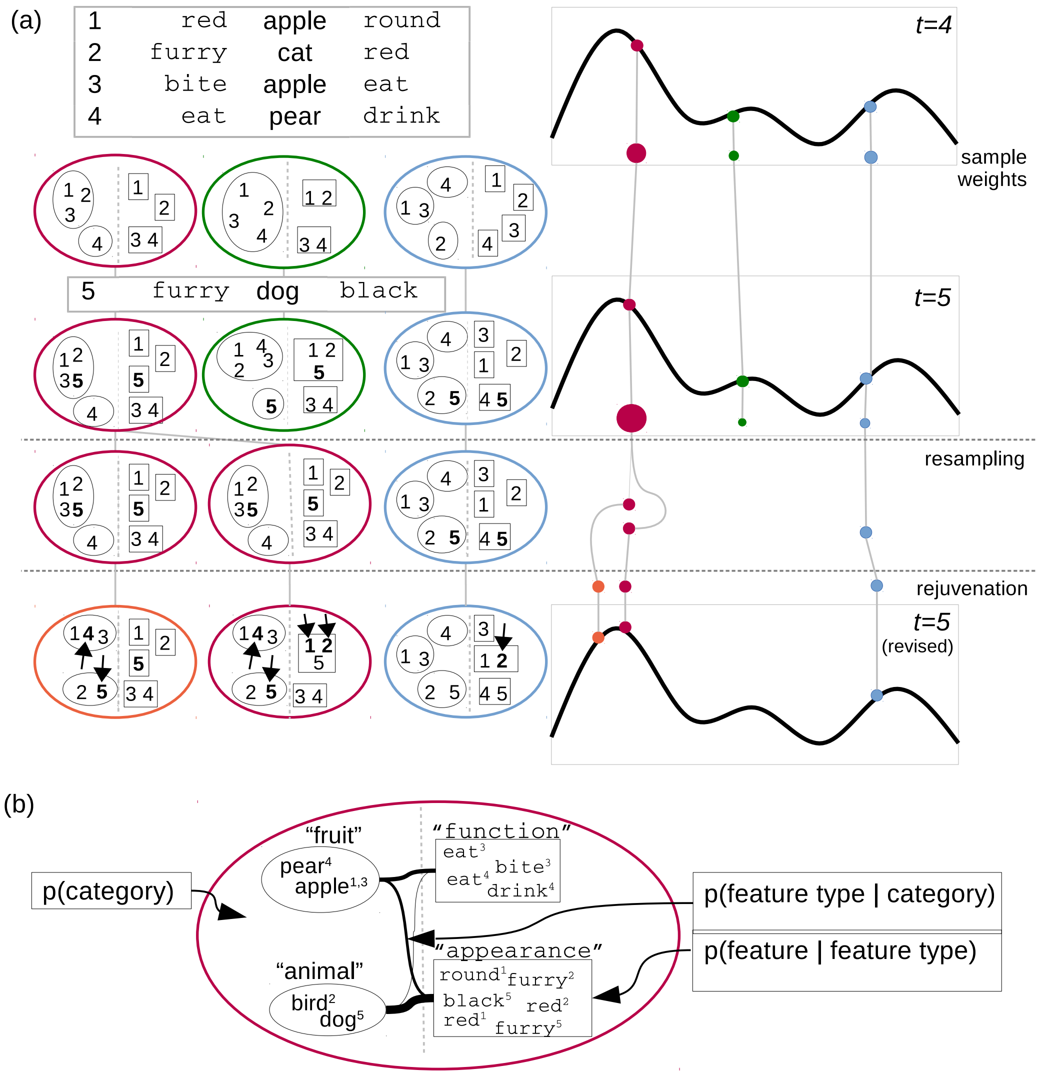


Figure 5.3: (a) Visualization of the particle filtering procedure for BCF with a 3-particle filter. Each particle corresponds to a clustering of the observed stimuli (category-assignments of observed concepts (circles), and feature type assignments of observed stimuli (boxes)) up to time *t* (left) . The collection of weighted particles is the current approximation of the posterior distribution over clusterings (right). The 5 stimuli observed by the filter are shown in the tables. We show one update step for all particles with stimulus 5, and one subsequent resampling and rejuvenation step. In the resampling step the highest-weight (red) particle is duplicated, replacing the lowest-weight (green) particle. In the rejuvenation step each particle revisits previous categorization decisions; (b) a zoom into the red particle at time $t = 5$ (revised). Each particle consists of a set of categories (left), category-specific distributions over feature types (indicated through weighted connections), and featuretype-specific distributions over features (right). We labeled the categories and feature types for illustration.

Algorithm 5 The Particle Filter for the BCF model.

```
Initialize particles by randomly partitioning first $d$ stimuli                ▷ Initialization
Initialize weights $\mathbf{w}^d = \frac{1}{N}$
for stimulus $t = [d+1..T]$ do
    for particle $n = [1...N]$ do
        $z_n^t = \{g_n^{d_t}, k_n^{c^{d_t}}\} \sim q(z_n^{1:t-1}|y^{1:t-1})q(z_n^t|z_n^{t-1}, y^t)$    ▷ Particle Update
                 $= p(g^{d_t} = i, k^{c^{d_t}} = j|\mathbf{f}^-, \mathbf{g}^-, \mathbf{k}^-; \alpha, \beta, \gamma)$   Equation (5.7)
        $S_n^t \leftarrow (S_n^{t-1}, z_n^t)$
        $\tilde{w}_n^t = w_n^{t-1} \times p(y^t|z^{t-1})$                                    ▷ Weight Update
             $= w_n^{t-1} \times \sum_i \sum_j p(g^{d_t} = i, k^{c^{d_t}} = j|\mathbf{f}^-, \mathbf{g}^-, \mathbf{k}^-; \alpha, \beta, \gamma)$
    $\mathbf{w}^t \leftarrow normalize(\tilde{\mathbf{w}}^t)$
    if $ESS(\mathbf{w}^t) \leq thresh$ then                                          ▷ Resampling
        $\mathcal{P}(i) \leftarrow \{\text{Mult}(\mathbf{w}^t)\}_{i=1}^N$
        $\mathbf{w}^t = \frac{1}{N}$
        for particle $n \in \mathcal{P}(i)$ do                                       ▷ Rejuvenation
            for rejuvenation point $o = [1...O]$ do
                $r \sim \text{unif}(0,1) \begin{cases} o \sim \text{unif}(1...t) \;;\; g^{d^o} \sim \text{eqn (5.3)} & \text{if } r \leq 0.8 \\ o \sim \text{unif}(1...C) \;;\; k^o \sim \text{eqn (5.5)} & \text{otherwise} \end{cases}$
```

each particle at any time. An algorithmic overview of the particle filter is displayed in Algorithm 5.

The particle filter propagates a set of weighted hypotheses or particles through time. We introduce a notion of time into our model by defining each individual stimulus observation as a time step. At any time t each particle corresponds to a categorization, a set of feature types, and category-feature type associations based on all stimuli observed up to time t, as illustrated in Figure 5.3 (b). At each time the current observation is integrated individually into each particle, and the particle state and its weight are updated with the new information and propagated to time $t+1$. This update process is schematically illustrated in Figure 5.3 (a), where stimulus 5 is integrated into each of the three particles containing a category and feature representation of previously encountered stimuli 1–4.

Technically, the particle filter for the BCF model is based on the technique of sequential importance sampling (SIS; see Section 3.3.3 for a technical introduction) and its structure resembles the structure of the particle filter developed for the BayesCat model in Section 4.3.2.1. Samples representing the target distribution of interest (here the posterior distribution of the BCF model) are obtained from an importance distribution (which is easier to sample from than the exact posterior distribution). Each sample is assigned a weight according to its discrepancy from the true target distribution. The importance distribution in our particle filter is represented through the set of particles available from the previous time step (approximating the target posterior at time $t-1$). Samples and weights are recursively updated from time $t-1$ to time t with information extracted from novel input stimuli observed at time t.

In each update step, we sample a feature type for stimulus d_t, g^{d_t}, as well as a category for stimulus d_t's concept, $k^{g^{d_t}}$, and update the relevant distributions as follows. Independently for each particle, we sample $\left(g^{d_t}, k^{c^{d_t}}\right)$ jointly from its posterior distribution conditioned on the particle's state from time $t-1$ which incorporates all information encountered up to this time (see lines 5–6 in Algorithm 5),

$$\begin{aligned} &P\left(g^d = i, k^{c^{d_t}} = j \middle| \mathbf{f}^-, \mathbf{g}^-, \mathbf{k}^-; \alpha, \beta, \gamma\right) \propto \\ &\qquad P\left(k^{c^{d_t}} = j \middle| \mathbf{g}_{k^\ell}, \mathbf{k}^-, \alpha, \beta\right) \quad \times \quad P\left(g^d_{k^{c^d}} = i \middle| \mathbf{g}^{-d}_{k^{c^d}}, \mathbf{f}^-, k^{c^d} = j, \beta, \gamma\right). \end{aligned} \tag{5.7}$$

The two components on the right-hand side correspond to equations (5.5) and (5.3), respectively. This posterior distribution takes into account both the prior probabilities over (k, g), represented through the particle swarm at time t, as well as the data in the observation at time t.

From an importance sampling perspective, this posterior distribution is *locally* optimal[3] in the sense that it minimizes the divergence of the importance weights (see Section 3.3.3, page 51 f. for a detailed discussion). The importance weights themselves are updated according to the predictive probability of observation t, and normalized to sum to one (see lines 6–7 in Algorithm 5).

Despite the local optimality of our importance distribution, the repeated approximate particle updates ultimately lead to divergence of the particle weights, resulting in few or even only one particle accumulating the majority of weight mass. Practically the empirical estimation of the posterior distribution through the set of weighted particles

[3] It is *locally* optimal because it assumes that previous particle states (i.e., variable assignments) are fixed.

then corresponds to a point estimate: the filter does not accurately represent the area under the target distribution. Figure 5.3 (a) illustrates the phenomenon of high variance in particle weights after particle propagation from $t = 4$ to $t = 5$ (right).

Resampling Like in the particle filter for the BayesCat model (Section 4.3.2.1) we alleviate this problem through *resampling*, a technique to improve the coverage of the sample space through our particles. A resampling step is executed whenever the variance among particle weights exceeds as threshold. Weight variance is measured through the *e*ffective sample size (ESS):

$$ESS(\mathbf{w}^t) = \left(\frac{1}{\sum_n (w_n^t)^2} \right). \tag{5.8}$$

Whenever this value falls below a threshold we resample N particles from a Multinomial distribution parameterized by the current set of particle weights N times with replacement. This leads to multiple copies of high-weight 'good' particles which represent the high-probability areas of the sample space, and which are further explored in the learning process. Low-weight ('bad') particles which do not get sampled in this process are discarded. After resampling, the particle weights are re-set to uniform since the resulting sample of particles corresponds to an empirical estimate of the previous weight distributions. The resampling process corresponds to lines 8–10 in Algorithm 5, and illustrated in Figure 5.3 (a).

Rejuvenation While resampling re-configures the particle filter such that it explores high-probability areas of the sample space, it impoverishes the sample: The representation of the distribution contains multiple copies of identical samples. Furthermore, low-weight particles are discarded which may well be only locally unlikely and might be accurate at a later stage after more data has been observed. To introduce diversity back into the resampled particle set a technique called *r*ejuvenation can be applied, which 'jiggles' each resampled particle without altering the distribution it represents (Gilks and Berzuini 2001, see Section s3.3.3.3, page 54).

We jiggle each resampled particle according to a Gibbs kernel which leaves the target posterior distribution unchanged. We randomly resample a fixed number of stimulus-feature type assignments to previously encountered stimuli, and concept-category assignments to previously encountered concepts. We resample feature types 80% of the time since these variables are assigned on the stimulus-level and whereas categories are assigned to concepts which are much smaller in number. The rejuvenation procedure is illustrated in the bottom part of Figure 5.3 (a), and technically summarized in Algorithm 5 (line 12 onwards). As discussed in Section 4.3.2.1 rejuvenation is also plausible from a cognitive point of view, since human learning is not strictly incremental: it is possible to re-consider knowledge in the light of newly gained observations or evidence.

5.3 Experiment 3: Large-Scale Category and Feature Learning

In this and the following section we evaluate model performance under the two inference algorithms. This section focuses on evaluating the quality of categories and feature types obtained by our model when applied to a large-scale corpus and learned in an optimal batch fashion with the Gibbs sampler, and compares BCF to a competitive text-based feature extraction model. We evaluate BCF as a *cognitive* model of human category and feature learning in Section 5.4. We present a detailed analysis of the categories, feature types and category-feature type associations. We start by presenting our data set, before we introduce the models used for comparison with our approach, and explain how system output was evaluated. We then report results on a series of experiments.

Data Like in the BayesCat evaluation (Chapter 4), our experiments are based on a set of basic level target concepts (e.g., *cat* or *chair*) from two norming studies (McRae et al., 2005; Vinson and Vigliocco, 2008), which were subsequently classified into 41 categories (Fountain and Lapata, 2010). The data set was described in detail in Section 4.4.1 (p. 80). Here, we use 34 of these categories as a goldstandard in our categorization experiments (comprising 492 concepts in total). We filter the set of categories for the joint category and feature evaluation, excluding very general categories such as THING or STRUCTURE. This decision is based on the intuition that it is difficult to identify characteristic features for them. As a heuristic, concepts were excluded if they were close to the root of WordNet (Fellbaum, 1998b), e.g., at depth 2 or 4.

To obtain the input stimuli for the BCF model, we used a subset of the Wackypedia corpus (Baroni et al., 2009), an automatically extracted and part of speech tagged dump of the English Wikipedia. For each target concept, we identified one corresponding article in Wackypedia. We identify mentions of target concepts c in the resulting data set and define context as the set of words making up the sentence c occurs in (except c). Next, we extracted a set of stimuli which consists of (a) every sentence from the concept's corresponding article, and (b) any sentence in a different article which mentions the concept. This resulted in a data set of 63,076 stimuli which we split into 60% training, 20% development and 20% test. We removed stopwords as well as words with a part of speech other than noun, verb, and adjective. Furthermore, we discarded words with an age of acquisition above 10 years (Kuperman et al., 2012) to restrict the vocabulary to frequent and generally familiar words, which reduced the set of context word types from 25,100 to 6,500 (by 74%).

Models and Parameters We compared the performance of BCF against BayesCat, our Bayesian model of category acquisition introduced in Chapter 4 and Strudel, a pattern-based model which extracts concept features from text (Baroni et al., 2010). In all experiments in this section BCF is trained in a batch fashion using the Gibbs sampler introduced in Section 5.2.1.

BayesCat induces categories which are represented through a distribution over target concepts, and a distribution over features (i.e., individual context words). In contrast to BCF, it does not learn types of features. In addition, while BCF induces a hard assignment of concepts to categories, BayesCat learns a soft categorization. Soft assignments are converted into hard assignments by assigning each concept to its most probable category as described in Section 4.4 (equation (4.9), page 84). We ran BayesCat on the same input stimuli as BCF, with the following parameters: the number of categories was set to $K = 40$, and the hyperparameters to $\alpha = 0.7, \beta = 0.1, \gamma = 0.1$. For the BCF model, we used the same value for $K = 40$, the number of feature types was set to $G = 75$, and the hyperparameters to $\alpha = 0.5, \beta = 0.5$, and $\gamma = 0.1$. Parameters were tuned on the development set. For both models, we report results averaged over 10 Gibbs runs, each represented as the final sampler state after 1,000 iterations. We used annealing during learning which proved effective for avoiding local optima.

Strudel automatically extracts features for concepts from text collections following a pattern-based approach. It takes as input a part of speech-tagged corpus, a set of target concepts and a set of 15 hand-crafted rules. Rules encode general, but quite sophisticated linguistic patters which plausibly connect nouns to descriptive attributes (e.g., *extract an adjective as a property of a target concept mention if the adjective follows the mention, and the set of tokens in between contain some form of the verb 'to be'.* (Baroni, 2010)). Strudel obtains a large set of concept-feature pairs by scanning the context of every occurrence of a target concept in the input corpus, and extracting context words that are linked to the target concept by one of the rules. Each concept-feature pair is subsequently weighted with a log-likelihood ratio expressing the pair's strength of association. Baroni et al. (2010) show that the learnt representations can be used as a basis for various tasks such as typicality rating, categorization, or clustering of features into types. We obtained Strudel representations from the same Wackypedia corpus used for extracting the input stimuli for BCF and BayesCat. Note that Strudel, unlike the two Bayesian models, is not a cognitively motivated *acquisition* model, but an optimized system developed with the aim of obtaining the best possible features from data.

5.3.1 Quality of Learnt Categories

In our first experiment we evaluate the quality of the categories induced by the three models presented above. The models produce hard categorizations, however, the cognitive gold standard we use for evaluation (Fountain and Lapata, 2010) represents soft categories. We obtained a hard categorization by assigning members of multiple categories to their most typical category (typicality scores are provided with the data).[4]

Method BCF and BayesCat learn a set of categories which we can directly compare to the gold standard. For Strudel, we produce a categorization as follows: we represent each concept as a vector over features (obtained from Wackypedia), where each component corresponds to the concept-feature log-likelihood ratios provided by Strudel. Following Baroni et al. (2010), we then cluster the vectors using K-means and the Cluto toolkit.[5] As for the other models, we set the number of categories to $K = 40$.

Metrics We use the same metrics for category quality assessment as in the evaluation of our BayesCat model in Chapter 4. They are described in detail in Section 4.4 (pages 84 ff.). To assess the quality of the clusters produced by the models, we measure

[4] http://homepages.inf.ed.ac.uk/s0897549/data/.

[5] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

	hom	**com**	**v1**	**pur**	**col**	**pcf1**
Random	0.32	0.28	0.30	0.20	0.14	0.16
BCF	0.68	0.64	**0.66**	0.59	0.52	**0.55**
BayesCat	0.65	0.59	0.62	0.57	0.45	0.50
Strudel	0.70	0.62	**0.66**	0.61	0.48	0.54

Table 5.2: Model performance on the category induction task.

purity (*pur*; the extent to which each learnt cluster corresponds to a single gold class) as well as its inverse, collocation (*col*; the extent to which all items of a particular gold class are represented in a single learnt cluster). Both measures are based on set-overlap, and we also report their harmonic mean ($pcf1$; Lang and Lapata (2011)). In addition, we report the V-measure ($v1$; Rosenberg and Hirschberg (2007)) and its factors measuring the homogeneity of clusters (*hom*) and their completeness (*com*). The two factors intuitively correspond to purity and collocation, but are based on information-theoretic measures.

Results Our results are summarized in Table 5.2. In addition to model performance we report PCF1 and V-Measure for a random clustering baseline (cf., Section 4.4.1, page 86 for a discussion). The score reflects average cluster quality of 10 random assignments. Overall, the results reveal that BCF and Strudel perform almost identically, and both outperform BayesCat. BCF *learns* the categories from data, whereas for Strudel we construct the categories post-hoc after a highly informed feature extraction process (relying on syntactic patterns). It is therefore not surprising that Strudel performs well, and it is encouraging to see that BCF does too. Also, note that Strudel tends to learn clean clusters at the cost of recall, whereas the tradeoff is less extreme for BCF. This patterns is commonly observed with pattern-based approaches, like Strudel. It is tempting to attribute the superior performance of BCF compared to BayesCat to the advantage of a *joint* learning process for categories and features in knowledge-lean, cognitively motivated models. However, there is another factor to consider. While BCF and Strudel are constrained to assign each concept to only one category, BayesCat induces a soft categorization which is turned into a hard categorization in a post-learning step. While this setting allows for more flexibility, it also induces more uncertainty and results in categorizations which resemble the gold standard less closely compared to the two other models.

5.3.2 Quality of Learnt Features

We next investigate the quality of the features BCF learns by letting the model predict the right concept solely from a set of features. If a model has acquired informative features, they will be predictive of the unknown concept. Specifically, the model is presented with a set of previously unseen test stimuli with the target concept removed. For each stimulus, the model ranks all possible target concepts based on the features **f** (i.e., context words).

Method We compare the ranking performance of BCF, BayesCat, and Strudel, like in the category evaluation above. For the Bayesian models, we directly exploit the learnt distributions. For BCF, we compute the score of a target concept c given a set of features as:

$$Score(c|\mathbf{f}) = \sum_{g} P(g|c)P(\mathbf{f}|g). \tag{5.9}$$

Similarly, for BayesCat we compute the score of a concept c given a set of features as follows:

$$Score(c|\mathbf{f}) = \sum_{k} P(c|k)P(\mathbf{f}|k). \tag{5.10}$$

For Strudel, we rank concepts according to the cumulative log-likelihood ratio-based association score over all observed features for a particular concept c:

$$Score(c|\mathbf{f}) = \sum_{f \in \mathbf{f}} association(c, f). \tag{5.11}$$

Metrics We report precision at rank 1, 10, and 20. We also report the average rank assigned to the correct concept. All results are based on a random test set of 2,000 previously unseen stimuli. To control for the possibility that the models are learning a strong (yet trivial) correlation between target concepts and identical words occurring as features, we also report results on a modification of our test set where we remove any mention of the target concept from the context, if present (the $-$tgt condition).

Results Our results on the concept prediction task are shown in Table 5.3. Both Bayesian models (BCF and BayesCat) outperform Strudel across all metrics and conditions. Strudel's extraction algorithm, which relies on pre-defined patterns, might be

		pr@1	**pr@10**	**pr@20**	**avg rank**
BCF	full	**0.12**	**0.50**	0.63	56.1
	−tgt	0.09	0.40	0.53	78.5
BayesCat	full	0.11	0.49	**0.64**	**37.7**
	−tgt	0.09	0.39	0.53	52.4
Strudel	full	0.07	0.33	0.47	64.4
	−tgt	0.07	0.35	0.49	62.2

Table 5.3: Model performance on the concept prediction task in terms of precision at rank 1, 10, 20, and average rank assigned. −tgt refers to the condition where we remove context words which are identical to the target concept as opposed to using the full context.

too restrictive with respect to the set of features it extracts which as a result are not as discriminative as the features learnt by BCF and BayesCat, which are a priori unrestricted. BayesCat and BCF perform comparably given that they learn from exactly the same data and exploit local co-occurrence relations in similar ways. BayesCat produces better average rank scores than BCF, while achieving lower precision scores. This can be explained by the fact that BCF assigns low ranks to correct concepts more reliably than BayesCat. Figure 5.4 shows the relative cumulative frequencies of the ranks assigned by the three models. We display the top ranks 1 through 20 (out of 492). As can be seen, BCF performs slightly better than BayesCat. Pairwise differences between the systems are all statistically significant ($p \ll 0.01$); using a one-way ANOVA with post-hoc Tukey HSD test).

Note that performance decreases for both Bayesian models in the −tgt condition, i.e., when occurrences of the target concept are removed from the context. Strudel is less affected by this given its pattern-based learning mechanism which is not prone to associating target word types with themselves. However, repetitions are a natural phenomenon from a cognitive standpoint and it seems reasonable to consider multiple occurrences of a concept as a canonical feature of the learning environment.

Overall, the precision scores may seem low. However, bear in mind that the models rank a set of 492 target concepts. A random baseline would achieve a $pr@1$ of only 0.002%. In addition, the target concepts we are considering are by design highly confusable: They were selected so that they form categories and are thus bound to

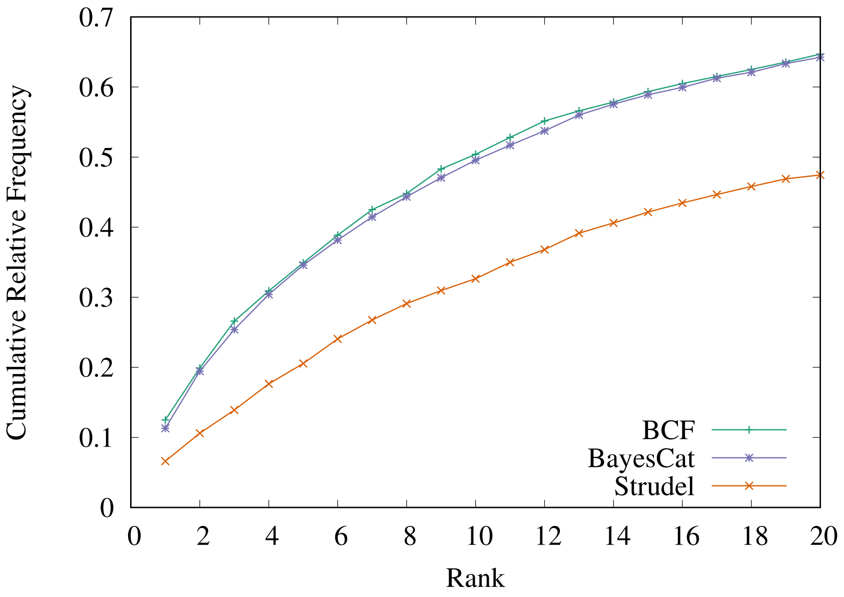


Figure 5.4: Number of times the correct target concept was placed within the top 20 ranks by BCF, BayesCat, and Strudel. The differences between the systems' performance are statistically significant (with the vector of 2K ranks as assigned by the particular systems as input; ANOVA (F=45.865 p< 0.01)).

share some features which makes the prediction task harder. Example output for all three models is shown in Figure 5.5. The models take context features "*journey move hundred mile strong*" and "*avoid cut quick claw tip*" as input and are expected to predict *salmon* and *finger*, respectively. Unlike Strudel, BCF and BayesCat rank *salmon* almost correctly and the other high ranked concepts are reasonable in the given context as well. For the second example, only Strudel predicts the correct concept correctly, but again the top-ranked concepts of the other two models are reasonable in the given context.

5.3.3 Quality of Learnt Feature Types

In this suite of experiments we evaluate two aspects of the feature types induced by our model: (1) Are they *relevant* to their associated category? and (2) Do they form a *coherent* class? Our evaluation followed the intrusion paradigm originally introduced to assess the output of topic models (Chang et al., 2009). We performed two intrusion studies using Amazon's Mechanical Turk crowd-sourcing platform.

In the feature intrusion study, participants were shown examples of categories and a list

salmon	`journey move hundred mile strong current reproduce`				
BCF	**salmon**	tuna	goldfish	lobster	fish
BayesCat	fish	radio	goldfish	**salmon**	clock
Strudel	train	house	apartment	ship	car

finger	`avoid cut quick claw tip painful`				
BCF	tent	ski	peg	curtain	hut
BayesCat	eye	ear	spider	leg	hair
Strudel	**finger**	toe	hair	tail	hand

Figure 5.5: Model output on the concept prediction task for *salmon* (top) and *finger* (bottom): the top part of each table shows the true concept (left) and the context provided to the model as input (right). The bottom part of the table shows the five most highly ranked concepts (left to right) for each model.

of feature types. Both categories and feature types were represented as word clusters. Example tasks are shown in Figure 5.6. One of the feature types was an ‘intruder’ not associated with the category in the model output, and participants were asked to detect the intruder feature type. If a model learns *relevant* feature types, we would expect participants to be able to identify the intruder relatively easily.

We also conducted a word intrusion study, where participants were shown a single feature type (again represented as a word cluster). One of the words, which was not highly associated with the feature type in the model output, was added as an ‘intruder’, Figure 5.7 displays two example tasks. Again, participants were asked to detect the intruder feature (i.e., word). If the features are overall *coherent* and meaningful, it should be relatively straightforward to identify the intruder.

Method We compared the feature types learnt by BCF and Strudel. We omitted BayesCat from this evaluation as it does not naturally produce feature types, rather it associates unstructured lists of features with categories. As mentioned earlier, Strudel does not induce feature types either, however, it associates concepts with features which can be post-processed to obtain feature types as follows. Given a category induced by Strudel (as explained in Section 5.3.1), we collected the features associated with at least half of the concepts in the category with a log likelihood score no less

'Select intruder feature type (right) wrt category (left).'	
ant hornet butterfly moth flea beetle grasshopper wasp caterpillar cockroach	○ `egg female food young bird`
	○ `ant insect butterfly wasp larva`
	○ `body air fish blood muscle`
	○ `sound human nerve bird brain`
	● `wear cover veil woman coat`
	○ `culture symbol popular feature animal`

veil coat hair fur glove cape hat cap bouquet scarf slipper	○ `wear cover veil woman coat`
	○ `white black color brown dark`
	● `cat box object litter mark`
	○ `eye tooth ear skin lip`
	○ `wear suit trouser woman garment`
	○ `animal feather material wool skin`

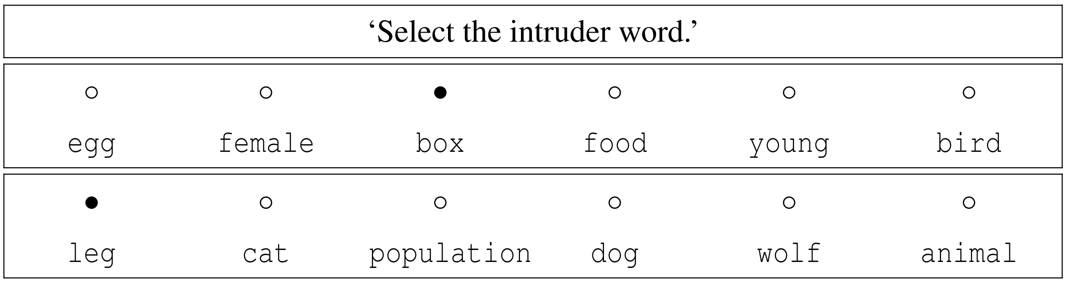
Figure 5.6: Two illustrations of the feature type intrusion task, with annotator instructions shown at the top. The correct responses are marked with a filled circle.

than 19.51.[6] We then clustered these features with K-means (using the Cluto toolkit) into $K = 5$ feature types. Note that the Strudel feature types were (i) elicited through a pipelined procedure, and (ii) are not shared across categories, but optimized for each category individually. We therefore expect Strudel to perform well in this evaluation.

For BCF, for each category k, we select the five feature types g with highest association $P(g|k)$, together with one intruder feature type g' which is highly associated with some other category k' but not with k. For Strudel we took the five feature types elicited through the procedure described above, and one random feature type from the global set of feature types. Each feature type was represented by a cluster of five words.

In the word intrusion task, participants were only shown feature types (i.e., word clusters) irrespectively of the associated category. BCF feature types g were represented as the set of the five words w with highest probability $P(f|g)$. In addition, we added one intruder word which had low probability under g but high probability under some other feature type. For Strudel, we represented feature types as a random subset of five words, and added an additional intruder word from the global set of features.

[6] Following Baroni et al. (2010), this number corresponds to a probability of co-occurrence below 0.00001, assuming independence.



‘Select the intruder word.’

○	○	●	○	○	○
egg	female	box	food	young	bird

●	○	○	○	○	○
leg	cat	population	dog	wolf	animal

Figure 5.7: Two illustrations of the word intrusion task, with annotator instructions shown at the top. The correct responses are marked with a filled circle.

For the feature type intrusion task, we evaluated a total of 40 categories for each model. Each participant assessed 10 categories per session (5 per model). Categories and feature types were presented in random order. For the word intrusion task, we evaluated a total of 66 feature types for each model. Participants saw 11 feature types per session, in randomized order. In both cases, we collected 10 responses per item. The instructions given to participants in the Mechanical Turk experiments are included in Appendix B.1 (for the feature type intrusion task), and in Appendix B.2 (for the word intrusion task). The full set of stimuli for both tasks and systems is available at `http://frermann.de/mturk2015/`.

Results We evaluated feature type relevance and coherence by measuring precision (the proportion of intruders identified correctly). We also use the Kappa coefficient to measure inter-subject agreement (Fleiss, 1981) on our two tasks.

Our results are presented in Table 5.4. Participants identify the intruder feature type correctly more than 50% of the time. The performance of Strudel is slightly better compared to BCF, both in terms of accuracy and Kappa (however the differences are not statistically significant, using a t-test). Again this is not surprising considering that Strudel's feature types were elicited through a highly informed, pipelined process. The results show that the simpler and cognitively plausible BCF model learns feature types of a quality comparable to a highly engineered, competitive system. Examples of feature types discovered by BCF and Strudel are shown in Figure 5.8, for the category CLOTHING. As can be seen, Strudel obtains a large number of action-related features (e.g., *replace, change, steal*). BCF creates more varied feature types. For example, the second cluster refers to external properties (e.g., color), the fourth cluster denotes related concepts such as hyponyms, and the last cluster contains CLOTHING materials.

	Feature Type Intrusion		Word Intrusion	
	Precision	**Kappa**	**Precision**	**Kappa**
BCF	0.52	0.23	**0.78**	**0.60**
Strudel	**0.56**	**0.26**	0.36	0.21

Table 5.4: Performance of Strudel and BCF on the feature type and word intrusion tasks. We report precision and inter-subject agreement (Fleiss' Kappa; all Kappa values are statistically significant at $p \ll 0.05$).

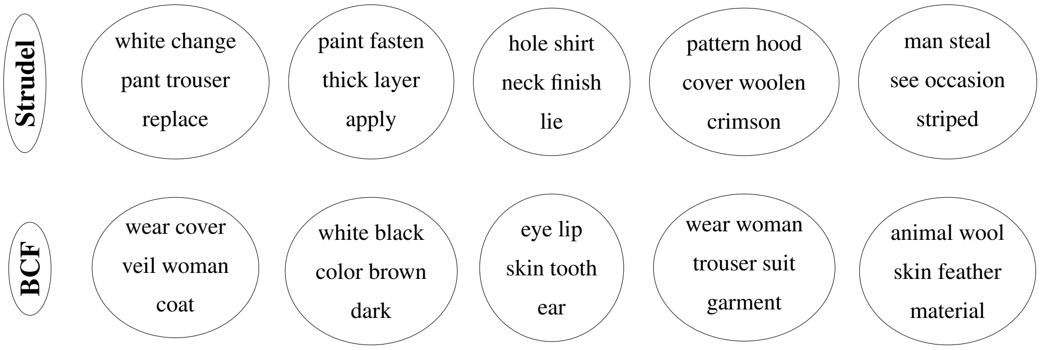

Strudel
white change pant trouser replace
paint fasten thick layer apply
hole shirt neck finish lie
pattern hood cover woolen crimson
man steal see occasion striped
BCF
wear cover veil woman coat
white black color brown dark
eye lip skin tooth ear
wear woman trouser suit garment
animal wool skin feather material

Figure 5.8: Example feature types learnt for the category CLOTHING by Strudel (top) and BCF (bottom).

Concerning the word intrusion task, we observe that participants are able to detect the intruder more accurately when presented with BCF feature types as compared to Strudel feature types (differences between Strudel and BCF are statistically significant at $p \ll 0.05$, again using a t-test). Figure 5.9 schematically illustrates the distribution over the number of annotators that agree on the correct intruder word for both Strudel and BCF. We can see (considering the combined green bars 9 and 10) that for more than 50% of the test items either 9 or 10 out of 10 annotators agreed on the correct intruder when presented with output from the BCF model. The results suggest that the feature types learnt by BCF are more coherent, and indeed express meaningful properties shared by concepts belonging to the same category. While being relevant to the category, Strudel's feature types do not seem to exhibit internal coherence to a similar extent. The examples in Figure 5.8 qualitatively confirm this result: It is more difficult to assign a meaningful label to the feature types induced by Strudel (top) than to those induced by BCF (bottom). For example, the second BCF feature type from the left could be labeled `color` and the rightmost one `material`. The mutual dependence of category formation and feature learning allows BCF to learn feature types which are

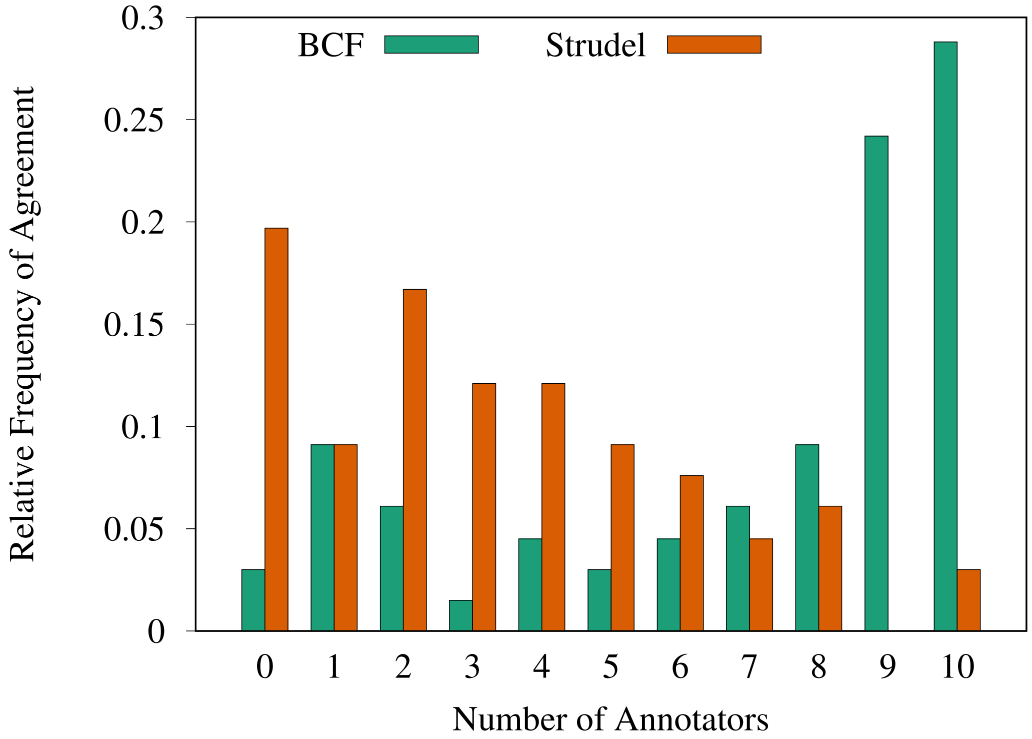


Figure 5.9: Illustration of Mechanical Turk annotator responses for the word intrusion task. Each bar shows the proportion of all responses in which $N \in \{1,...,10\}$ annotators agree on the correct label for BCF (green) and Strudel (orange).

both relevant and individually interpretable.

5.3.4 Discussion

We applied our joint model of category and feature learning, BCF, to large-scale encyclopedic text extracted from Wikipedia, and showed that it effectively captures category- and associated structured featural information encoded in this data. Evaluation of the inferred categories and their features shows that BCF performs competitively compared to a system specifically engineered to extract high quality features, despite the more complex learning objective, and the knowledge-lean approach.

We approximated the learning environment with large text corpora extracted from Wikipedia. However, we do not claim that the induced features closely correspond to features produced by humans in human feature elicitation studies. Instead, we show, through crowdsourcing-based human evaluation, that the learnt features are meaningful in that they are relevant to their associated category and form a coherent class.

Having demonstrated how our model performs on a broad task and under an optimal, batch learning algorithm, the following section will focus on BCF as a cognitive model of human learning.

5.4 Experiment 4: Child Category and Feature Learning

In this section we investigate the learning process of the BCF model under conditions which more closely resemble those faced by children acquiring categories and associated features. We apply BCF to a corpus of child-directed language, approximating the learning environment that children are exposed to. We train BCF with the incremental learning algorithm introduced in Section 5.2.2, which approximates the target posterior distribution in an sequentially performing one sweep over the training data and recursively improving its estimate. We refer to this incremental version as i-BCF (and to its batch version learnt with a Gibbs sampler as BCF).

Our evaluation is structured into two parts. In the first part we are interested in verifying that i-BCF induces meaningful categories and feature types. To this end we compare the incremental i-BCF to BCF, its Gibbs sampling-based counterpart. Both models are trained on a corpus of child-directed language. We quantitatively evaluate the induced categories, as well as the learnt feature representations. We also present qualitative examples of categories and feature types induced by the particle filter. Having confirmed that the incremental i-BCF model is capable of learning meaningful categories and representations, we move on to investigate the effect of resource constraints on the learning process: we compare the performance of i-BCF models trained with particle filters with varying numbers of particles.

Data The child-directed speech corpus underlying our BayesCat experiments (Section 4.5) is not appropriate as input data to the (i-)BCF models, because the respective stimuli comprise a context window of only ±2 words. Each concept mention is thus represented by a very restricted feature set, which is likely too limited for learning structured feature representations.[7] Instead, we extracted a dense longitudinal corpus

[7] We confirmed this hypothesis experimentally: i-BCF models trained on the CHILDES corpus used in Section 4.5 did not learn meaningful features and showed less discernible learning curves.

of child-directed speech from the CHILDES database (MacWhinney, 2000), comprising frequent recordings of child-parent interactions over an extended time span.[8] From this underlying data set we extract stimuli from it with a slightly larger context window of ± 3 words.[9]

We create an input corpus, comprising four sub-corpora, all of which contain transcribed speech data from natural interactions of children with their caretakers (mostly their mothers) at home:

- The 'Thomas' corpus (Lieven et al., 2009) contains data from interactions with one monolingual British English child who was recorded over a period of 3 years (from age 2 to age 5). Recordings were made five times a week for one hour during the first year, and for one hour per month in the two following years.
- The 'MPI-EVA-Manchester' corpus (Theakston et al., 2015) contains recordings of interactions from two monolingual British English children. Their interactions with caretakers were recorded between age 2 and 3. Recordings were made 10 times per month most of the time, but more frequently (10 times a week) for the first and last two months of the recording period.
- The 'Manchester' corpus (Theakston et al., 2001) comprises recordings of 12 monolingual British children between 2 and 3 years old. The recordings are less dense with two recorded sessions (30 minutes each) in every 3-week period for one year.
- The 'Providence' corpus (Demuth et al., 2006) contains recordings form longitudinal studies of 6 monolingual American English children aged between one and three years. Recordings were made for one hour every two weeks.

We filter all child-produced utterances from the corpora, so that our input corpus consists only of child-directed language. We divide the documents into time-stamped subsets by conflating documents by the age of the child being spoken to into bins covering one month. The earliest covered time period is 0 years 11 months, and the last period is 4 years 11 months. We remove stop words, comprising a standard list of function

[8] This corpus also underlies the cognitive experiments on meaning development in child concept representations in Chapter 6, and was thus constructed with a demand of temporally dense recordings in mind.

[9] Note that spoken, child-directed language is largely made up of short utterances which frequently switch topic, so that even larger context windows will result in a high proportion of irrelevant information in individual stimuli.

words, names, as well as a list of non-content words specific to child-directed speech (for example different forms of ‘mum’ and ‘dad’).

We extracted input stimuli which consist of one target concept within a ± 3 word context window. We restrict the set of context words, removing all words which occur fewer than 100 times in the resulting corpus of stimuli, leading to a reduction of context word types from 10,922 to 1,459 (by 87%). We then only keep input stimuli which, after context filtering, still possess their full context (i.e., six surrounding words). The resulting corpus covers 119 target concepts from 26 different categories. This is a subset of the 492 concepts used in the Wikipedia-based evaluation in Section 5.3. Due to the nature of child-directed speech the remaining concepts did not appear in our final corpus. We extract a total of $47,639$ stimuli comprising $1,459$ context feature types. During the incremental learning process, the extracted stimuli are presented to the model in chronological order, sorted with respect to the age of the addressed child.

Method We train i-BCF models on the CHILDES corpus described above using particle filters with varying numbers of particles, $N \in \{1,5,10,20,50,100\}$. We set the effective sample size threshold for resampling to $ESS(\mathbf{w}) = 0.5 * N$ and rejuvenate 100 previously observed stimuli after each resampling process. These parameters are identical to the settings used in the incremental BayesCat experiment in Chapter 4. The i-BCF parameters are set to the following values: the number of categories $K = 30$, the number of feature types $G = 35$, and the hyperparameters $\alpha = 0.3, \beta = 0.1, \gamma = 0.1$. The parameter values were adapted to the smaller nature of the CHILDES corpus as compared to the Wikipedia corpus, but not tuned exhaustively. For the particle filters with $N > 1$ particles we report performance as the score of the best-performing particle. All reported quantitative results are based on averages over 10 runs of any N-particle filter.

In order to contextualize the performance of the i-BCF models, we also train a BCF model (using a Gibbs sampler) on the CHILDES corpus. The BCF parameters are set to the same values as those of i-BCF, and we ran the Gibbs sampler for 1,000 iterations. Like in the Wikipedia experiments, we use annealing in order to avoid local optima. Again, all results are averages over 10 runs of the sampler. Unlike in the Wikipedia experiments, we do not report results on Strudel (Baroni et al., 2010) here. Strudel’s feature extraction mechanism relies on a set of syntactic rules, which were defined for grammatical (written) language. The language of the CHILDES corpus, however, is

	hom	**com**	**v1**	**pur**	**col**	**pcf1**
Random	0.54	0.48	0.51	0.35	0.29	0.32
BCF	0.72	0.67	0.69	0.62	0.55	0.58
i-BCF	0.47	0.45	0.46	0.56	0.56	0.56
BCF (Wikipedia)	0.68	0.64	0.66	0.59	0.52	0.55

Table 5.5: Quality of categories induced by the i-BCF model (with 100 particles) and the BCF model when trained on the CHILDES corpus. We also report a random clustering baseline (Random). For comparison we repeat the BCF results on the Wikipedia corpus (note that due to differences in the underlying test set, the scores on the CHILDES and the Wikipedia corpus are not directly comparable).

spoken and child-directed and hence frequently non-standard and ungrammatical.

We report categorization quality in terms of two automatic clustering evaluation scores, purity, collocation, pcf1, and V-measure, as described in Section 5.3.1. In addition, we compare i-BCF against BCF on its ability to predict a missing target concept based on the stimulus context, as in the evaluation described in Section 5.3.2. We use a random selection of 300 unseen test stimuli in this evaluation, and report precision at ranks 1, 10 and 20. In addition to these task-based evaluation metrics we also report the learning curves in terms of model log-likelihood i-BCF models with varying numbers of particles.

5.4.1 Quality of Learnt Categories and Features

We compared the output of the batch BCF model against the categories and features learned by the incremental i-BCF model. All results reported in this section are taken from the final representations induced by the highest-weighted particle of a 100 particle filter after observation of the full training corpus, unless otherwise specified.

Table 5.5 compares the quality of induced categories of BCF and i-BCF, as well as a random baseline (Random). Details on the random baseline and its misleadingly high V-Measure scores are provided in Section 4.4.1 (page 86). We repeat the BCF categorization performance on the Wikipedia corpus from Section 5.3. The quality of categories induced by BCF on the two corpora are comparable, although the numbers do not align directly since the target concept data sets for the two corpora differ in

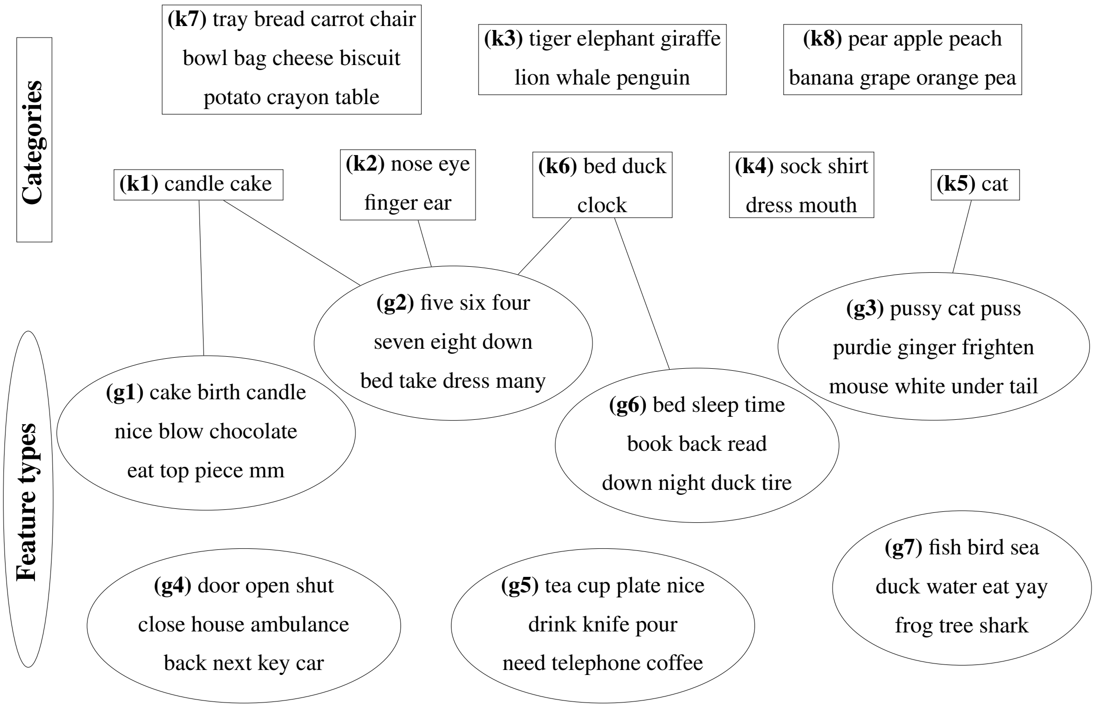

Categories
(k7) tray bread carrot chair bowl bag cheese biscuit potato crayon table
(k3) tiger elephant giraffe lion whale penguin
(k8) pear apple peach banana grape orange pea
(k1) candle cake
(k2) nose eye finger ear
(k6) bed duck clock
(k4) sock shirt dress mouth
(k5) cat
Feature types
(g2) five six four seven eight down bed take dress many
(g3) pussy cat puss purdie ginger frighten mouse white under tail
(g1) cake birth candle nice blow chocolate eat top piece mm
(g6) bed sleep time book back read down night duck tire
(g7) fish bird sea duck water eat yay frog tree shark
(g4) door open shut close house ambulance back next key car
(g5) tea cup plate nice drink knife pour need telephone coffee

Figure 5.10: Examples of categories (top) and feature types (bottom) inferred by the i-BCF model (with 100 particles) trained on the CHILDES corpus. Connecting lines indicate a strong association between the category and the respective feature type.

size. BCF clearly outperforms i-BCF which is unsurprising given its ideal batch learning behavior. This observation is also in line with our comparison of the incremental BayesCat model with its batch counterpart in the previous chapter (Table 4.6, Page 93).

Figure 5.10 displays qualitative examples of the categories and associated feature types induced by the 100 particle i-BCF model from the CHILDES corpus. The examples confirm that despite the quantitative gap in performance, the incremental model still learns discernible categories and meaningful featural associations from child-directed language. Categories such as ANIMAL (k3), BODYPART (k2) or FOOD and FRUIT (k7, k8) emerge. A `number` / `counting` related feature type emerges (g2) which is associated with a BIRTHDAY/CAKE category (k1), the BODYPART category (including the concept *finger*) (k2) as well as category k6 which includes the concept *clock*. Overall, the feature types are not as interpretable as those induced from the Wikipedia data (cf., Figure 5.5) which is unsurprising given the noisier data set of natural, child-directed speech which, in contrast to Wikipedia, is not constructed with the explicit single purpose of knowledge transfer.

	pr@1	**pr@10**	**pr@20**
BCF	0.12	0.51	0.67
i-BCF	0.07	0.32	0.49
Gibbs (Wikipedia)	0.12	0.50	0.63

Table 5.6: Comparison of the concept prediction performance of the i-BCF model (with 100 particles) and BCF when trained on the CHILDES corpus. For comparison we repeat the BCF results on the Wikipedia corpus (note that due to differences in the underlying test set, the scores on the CHILDES and the Wikipedia corpus are not directly comparable).

We compare the performance of BCF and i-BCF on the feature prediction task in Table 5.6. We report precision results at ranks 1, 10 and 20.[10] Models are presented with the context of an unseen input stimulus and predict a ranking of target concepts based on their probability in the given context, as described in Section 5.3.2. Again BCF's performance on Wikipedia is overall comparable to its performance on the CHILDES corpus (but note that the numbers do not align directly due to different model settings and test sets). Like in the previous evaluation, there is a drop in performance for i-BCF, when compared to BCF. Note that i-BCF still performs significantly above chance (random choice would lead to an expected precision at rank 1 score of 0.008).

Figure 5.11 lists examples of model output in the concept prediction task for i-BCF models with 1-particle ($N = 1$), and 100-particles ($N = 100$), and for BCF. All models were trained and tested on the CHILDES corpus. The first example shows model predictions for a test stimulus with context {silver vest red black car color}. Both the 100-particle i-BCF and BCF rank the correct concept (*car*) among their top 5 predictions, which are overall coherent and meaningful. The 1-particle i-BCF predictions are less relevant and less coherent. Examples 2–4 show instances where the correct concept is highly ranked by most systems. Examples 5 and 6 display examples where the advantage of batch BCF model becomes apparent. The context of example 6, {hair love night eat worm food}, leads BCF to correctly consider the latter features and predict a set of relevant animals within the top 5 ranks. The i-BCF models on the other hand seem go be deceived by the heterogeneous feature set of the stimulus, and make less consistent predictions.

[10] Unlike in Experiment 3, we do not report average rank results because no meaningful pattern emerged suggesting that they do not reflect model performance.

1	***car*** silver vest red black car color					(rank)
i-BCF $N = 1$	cat	pajamas	bucket	orange	shirt	(24)
i-BCF $N = 100$	train	**car**	helmet	bicycle	wheel	(2)
BCF	ambulance	bicycle	**car**	tractor	helmet	(3)

2	***crayon*** nicole chalk play move over draw					(rank)
i-BCF $N = 1$	**crayon**	crane	ball	envelope	fence	(1)
i-BCF $N = 100$	bed	**crayon**	hand	pot	ball	(2)
BCF	bus	ball	train	pie	**crayon**	(5)

3	***cup*** kettle boil ready tea lift orange					(rank)
i-BCF $N = 1$	knife	bin	bag	potato	bread	(6)
i-BCF $N = 100$	bed	crayon	hand	pot	**cup**	(5)
BCF	**cup**	bottle	orange	tray	fridge	(1)

4	***cat*** nice ginger pussy realize watch happy					(rank)
i-BCF $N = 1$	door	fridge	**cat**	key	gate	(3)
i-BCF $N = 100$	nose	ear	eye	**cat**	bear	(4)
BCF	**cat**	mouse	fence	tail	dog	(1)

5	***house*** actual park outside roof catch fire					(rank)
i-BCF $N = 1$	train	ambulance	garage	horse	bicycle	(21)
i-BCF $N = 100$	train	car	helmet	bicycle	wheel	(23)
BCF	telephone	**house**	mouse	door	spider	(2)

6	***bird*** hair love night eat worm food					(rank)
i-BCF $N = 1$	hand	cheese	hair	plate	toilet	(36)
i-BCF $N = 100$	bin	chair	box	hair	table	(22)
BCF	caterpillar	butterfly	frog	**bird**	crocodile	(4)

Figure 5.11: Model output for the concept prediction task for i-BCF with 1 particle ($N = 1$), i-BCF with with 100 particles ($N = 100$), and by the batch BCF model (BCF). The top row of each example shows the true concept to be predicted (bold italics; left) and the context provided to the model as input (right). The bottom part of each example shows the five most highly ranked concepts (left to right) for each model, as well as the rank of the correct concept in brackets on the right.

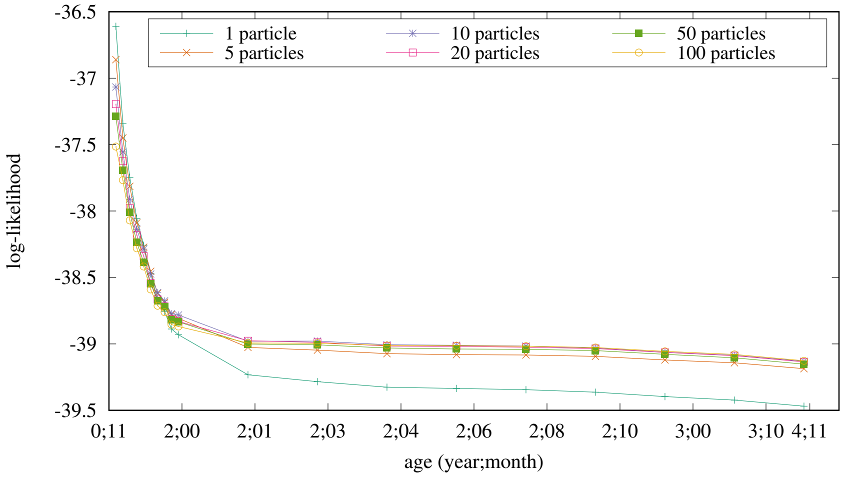


Figure 5.12: Model log-likelihood development on the CHILDES corpus for i-BCF models with varying numbers of particles.

The qualitative examples do not only reflect the quantitative results presented in Table 5.6, but also show that the task is far from trivial. Due to the removal of stopwords and low-frequency words, the local context of a concept may not be highly predictive of the missing concept. The 300 unseen test stimuli were selected at random without filtering for meaningful contexts.

5.4.2 Analysis of Memory Constraints

In this section we investigate i-BCF's incremental learning process itself. We are interested in (a) whether discernible learning curves in terms of continuous improvement in performance emerge; and (b) how this process is influenced by restricting the number of particles available to the particle filter. Increasing the numbers of particles allows for a better empirical estimate of the sample space at the cost of exceedingly high memory usage: each particle holds a sample from the posterior distribution of interest which is individually updated with newly observed information. Clearly human cognitive processing capabilities are limited by memory and we investigate the influence of the number of particles on the quality of the representations learnt by our i-BCF models.

Method i-BCF models are sequentially presented stimuli of child-directed language, chronologically sorted with respect to the addressed child. We compute learning curves for i-BCF models trained with $N \in \{1, 5, 10, 20, 50, 100\}$ particles for a variety of evaluation metrics. We report learning curves based on (1) model log-likelihood, a model-internal measure for convergence; (2) category quality in terms of purity, collocation, pcf1 measure, as well as homogeneity, completeness and v1; and (3) feature-based concept prediction in terms of rank precision scores.

Results Figure 5.12 displays the development of model log-likelihood for particle filters with varying numbers of particles. The overall log-likelihood values improve for i-BCF models with an increasing number of particles (higher is better). This is expected because more particles provide better coverage of the probability space and hence approximate the posterior distribution is approximated increasingly accurately. The difference in performance between the 1-particle filter and filters with multiple particles is most pronounced, whereas filters with multiple particles available perform very similarly. Note that performance differences are to some extent smoothed out by the fact that the learning curves are based on average values for 10 runs of the respective filters. Similar to our observation for the BayesCat model the log-likelihood flattens out towards the end of the learning curve (Figure 4.8b, page 100). While ideally it should eventually improve, we suspect that the size of the stimuli set used in this experiment was too small.

Figures 5.13 and 5.14 display the development of category quality in terms of purity, collocation and their harmonic mean (pcf1) in Figures 5.13a–5.13c, as well as homogeneity, completeness and their harmonic mean (V-measure) in Figures 5.14a–5.14c. Overall, clearly discernible learning curves in terms of continuously improving quality emerge which gives further indication that our i-BCF models learn categories and features effectively in a joint and incremental fashion. As a further overall pattern we can observe that more particles lead to higher-quality category estimates throughout the board.

Finally, Figure 5.15 displays the incremental process of learning to predict concepts based on their surrounding features, and thus provides a measure of the development of the quality of learnt featural representations over time. We report prediction precisions at ranks 1, 10 and 20. Overall we observe again improvement over time. i-BCF models with more particles again show superior performance, particularly in the early learning phase.

Comparing performance across evaluation metrics (log-likelihood in Figure 5.12, cat-

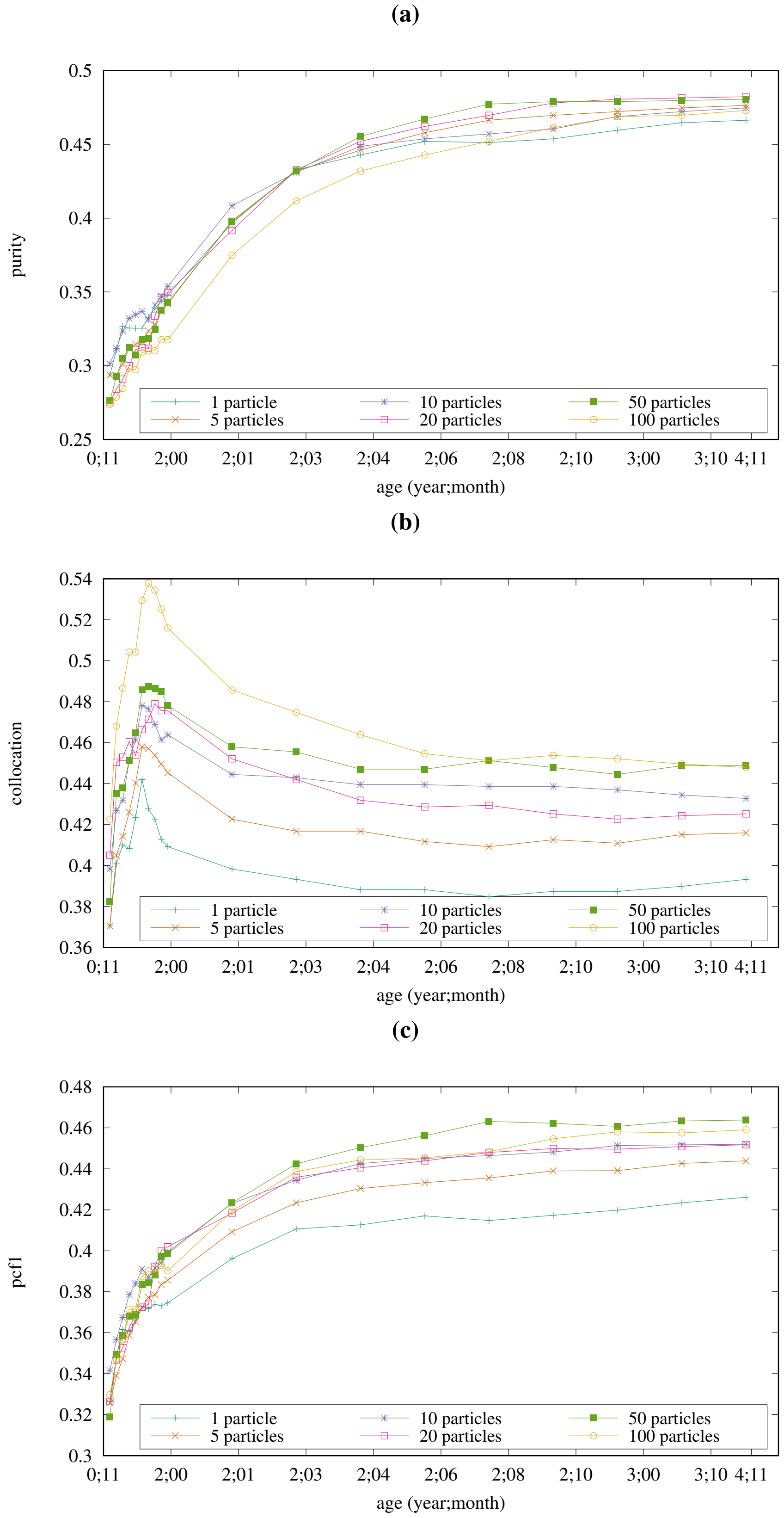
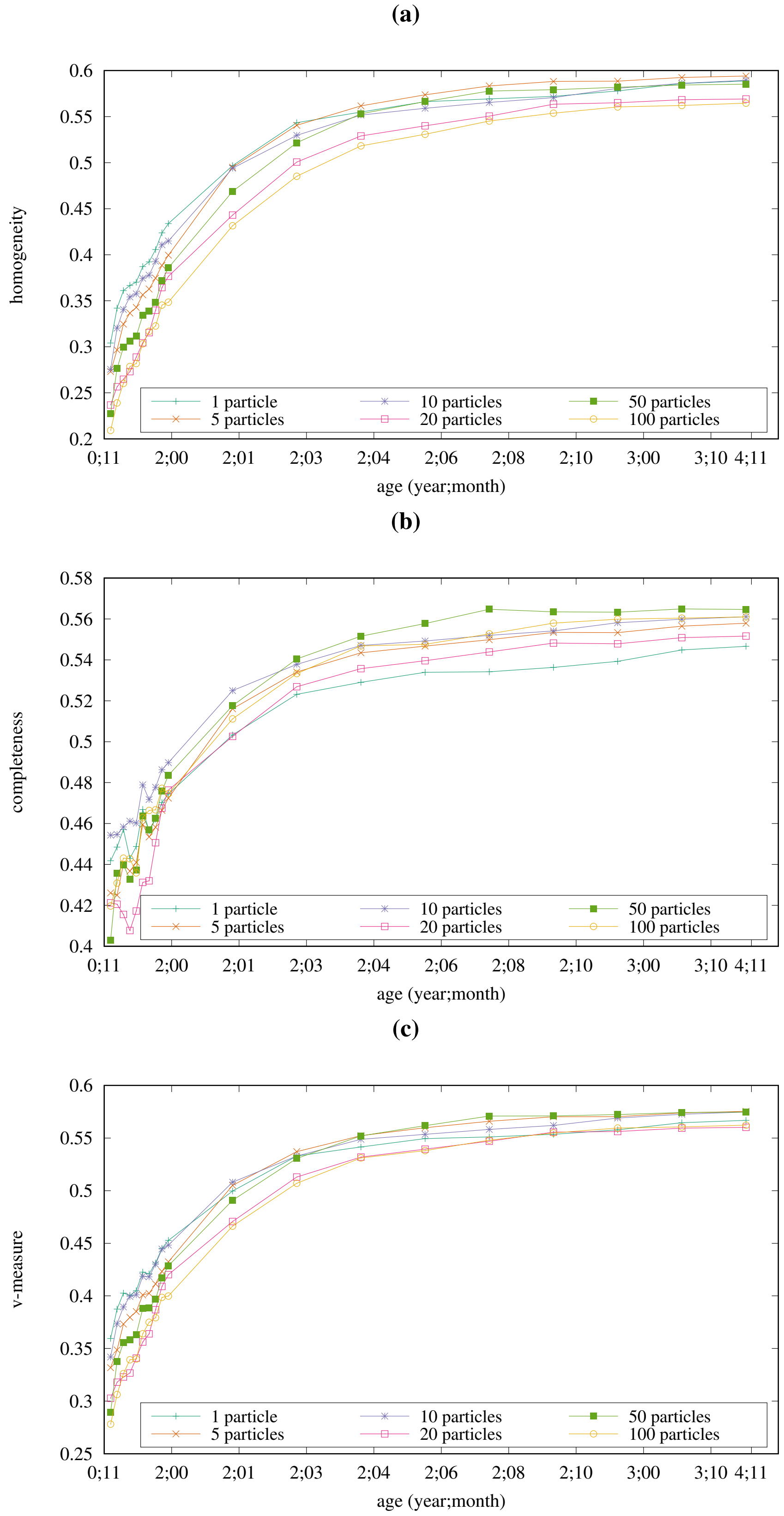


Figure 5.13: Purity (a), Collocation (b) and PCF1 (c) learning curves on the CHILDES corpus for i-BCF models with varying numbers of particles.


(a)

(b)

(c)

Figure 5.14: Homogeneity (a), Completeness (b) and V-Measure (c) learning curves on the CHILDES corpus for i-BCF models with varying numbers of particles.

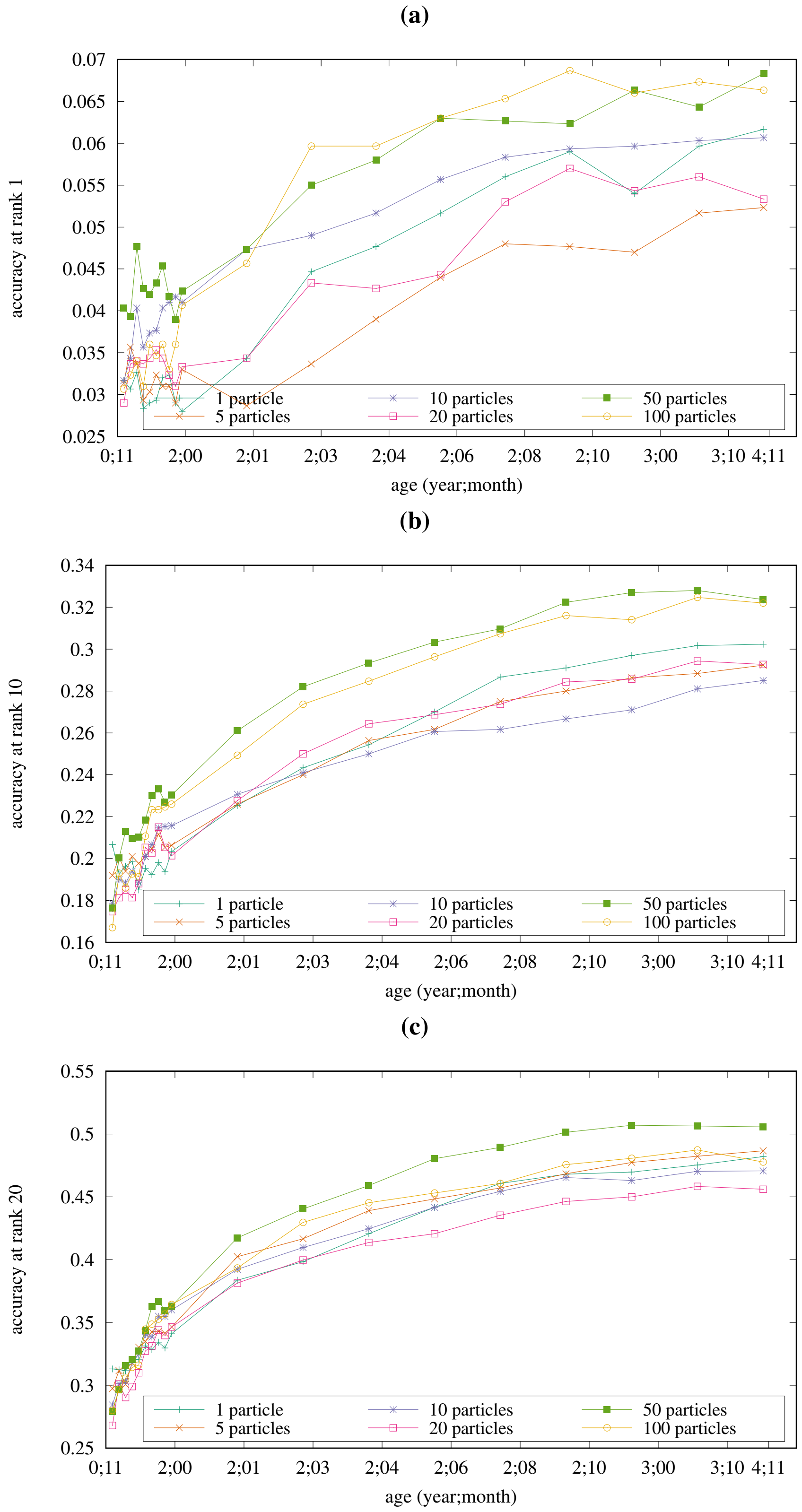


Figure 5.15: Concept prediction learning curves for accuracy at rank 1 (a), rank 10 (b), and rank 20 (c) on the CHILDES corpus for i-BCF models with varying numbers of particles.

egorization in Figures 5.13 and 5.14, and concept prediction in Figure 5.15) shows that the gap in performance is particularly pronounced for the 1-particle filter compared to filters with multiple particles, especially in terms of the model-internal log-likelihood metric. This is unsurprising as filters with more particles can explore the model space increasingly effectively and are more likely to cover high-probability regions of the parameter space. For the task-based evaluations, however, the performance across filters with varying number of particles is more comparable. This is in line with our findings in the context of the BayesCat model discussed in Section 4.5.2 (page 106 f.): Particle filters with a very moderate number of particles perform competitively across task-based evaluations. Viewed in the context of human learning, this is an encouraging result since it seems unlikely that humans have the cognitive capacities to maintain a large number of 'hypotheses' in parallel for any learning task. We showed that our incremental model can learn categories and featural representations effectively even under limited resources.

5.4.3 Discussion

Learning to group concepts into categories and to identify their relevant features is a formidable task children face. In this section we modeled the joint category and feature acquisition process with a cognitively motivated Bayesian model. We showed that our model induces discernible categories and featural representations from a corpus of natural, child-directed language and under an incremental learning algorithm which approximates the nature of human learning.

We presented our model with transcribed child-directed speech, approximating the environment from which a child acquires category knowledge.[11] The restriction to purely linguistic input does not faithfully capture the breadth of information a child has access to – visual and pragmatic cues are undoubtedly essential for any learning process. Nevertheless, our models induced meaningful categories and featural representations. Following the evaluation of BayesCat in Chapter 4, these results provide

[11] As discussed in Chapter 4 (page 107), high-frequency function words as well as rare words are filtered from the input corpora presented to our models. This preprocessing step is very common in modeling information from co-occurrence statistics in text corpora, and has been shown to increase the interpretability and relevance of induced information. The models presented in this thesis are sensitive to high-frequency words and induced features would likely be dominated bu function words. From a cognitive point of view filtering high and low-frequency term can also be interpreted as an approximation of attention: through pragmatic information such as prosody or gesture the child's attention is guided towards relevant words and their referents (Dominey and Dodane, 2004).

further support to our hypothesis that linguistic input is a rich source of information which incorporates much of the structure and information necessary for conceptual learning.

We captured the incremental nature of human learning in our models by using a sequential Monte Carlo learning algorithm (particle filter). We showed that, while the quality of learnt categories and representations decreases compared to an ideal batch learner, meaningful representations nevertheless emerge. The batch learner, a Gibbs sampler, can be viewed as an ideal observer which holds the complete training data set in memory and repeatedly iterates over the data to improve the learnt representations. A particle filter, much more like humans, is presented with training data points sequentially, observing one stimulus at a time, and immediately integrates the newly observed information into the knowledge extracted from previously seen input.

With every input stimulus, our particle filter samples (a) a feature type for the stimulus and (b) the category of the concept mentioned in the stimulus. Questioning category membership with every observation of a concept, however, seems exhaustive. A more realistic learner might only periodically re-consider its current belief about a categorization whenever the associated featural representations have been skewed by recent observations. One could imagine a resample-move based particle filter which samples stimulus-level feature types with every input stimulus, but resamples a categorization of concepts only periodically. However, we leave this project to future work.

5.5 Summary

This chapter presented BCF, a cognitively motivated Bayesian model which jointly learns categories and their features, arguing that the two tasks are co-dependent. We derived two approximate learning algorithms: a Gibbs sampler, which is an ‘ideal’ batch learner with access to all training data throughout the learning process; and an incremental learner in form of a particle filter, an instantiation of a sequential Monte Carlo algorithm which more faithfully resembles the incremental nature of *human* learning. We investigated the incremental learning procedure as well as the influence of memory constraints on the learning process through the particle filter.

Our model learns features from raw text without relying on elaborate pre- or post-processing or pre-defined knowledge (e.g., in terms of syntactic patterns). We showed that high quality categories, feature types and their associations emerge from large-scale encyclopedic data when estimated with a batch learner. In addition, we applied our model to a corpus of cognitive data of child-directed language, approximating the environment a child is exposed to when acquiring categories and their representation. Evaluation of the quality and development of acquired categories and features demonstrated our model's effectiveness under an incremental learning algorithm. We also showed that its performance degrades gracefully when resource constraints are imposed on the learning process.

An interesting direction for future work would be to learn feature types from multiple modalities (not only text) and to investigate how different information sources (e.g., visual or pragmatic input) influence feature learning. The BCF model learns descriptive feature types represented as a collection of feature values. In addition to such descriptive features (e.g., `behavior`) categories also possess *defining* features (e.g., `animate`) which are bound to one particular value. Extending the model in a way that allows to learn qualitatively different types of features is desirable from a cognitive perspective. In addition, it would be interesting to investigate the emergence of feature types with nonparametric Bayesian methods.

Another interesting avenue for future work would be to explicitly evaluate the *incremental formation* of categories and their featural representations experimentally against behavioral data obtained from children: do categories and features form in the same order as they do in child acquisition, and do the intermediate representations captured by our i-BCF model resemble those found in young children? We expect this evaluation to be challenging due to the difficulty to obtain longitudinal developmental data for children.

Finally, the BCF model can be applied to tasks beyond those discussed in this chapter. For example, one could learn definitions (aka features) of terms (aka concepts) in specialist fields (e.g., finance, law, medicine) or monitor how the meaning of words or concepts as represented by their features changes over time.

Chapters 4 and 5 of this thesis were concerned with modeling the acquisition of categories and features: we investigated the acquisition process of natural categories and their structured featural representations from large-scale naturalistic input data with computational cognitive models. Our evaluation implicitly assumed the existence of one true categorization and, consequently, featural representation of concepts. The success of a learner was measured by the extent to which its output resembles this gold standard categorization. Human conceptual knowledge, however, is flexible and susceptible to change: Conceptual representations have been shown to be *dynamic* and adapt over time and situations. The following chapter is dedicated to these phenomena and investigates the development of linguistic and conceptual representations over time.

Chapter 6

Modeling Meaning Change over Time

The previous two chapters investigated the process of category and feature acquisition assuming a single true "gold" categorization against which the quality of the model output was evaluated. This assumption is reasonable to the extent that there exists a strong agreement on the meaning of concepts and categories among members of a society in order to ensure effective communication. Various phenomena suggest, however, that conceptual representations can be *dynamic*, and flexibly adapt to the situation or environment. Concepts and categories are our mental tools for efficiently representing and interacting with the world. These representations must be necessarily flexible and able to adapt to the ever changing environment they represent. Gradual individual (e.g., through education) or societal (e.g., cultural or technological innovation) development over time triggers a smooth adaptation of concepts to match the demands of their users.

Concept representations change in the course of *learning*, and this phenomenon has been observed in both adults (Schyns and Rodet, 1997; Navarro et al., 2013) and children (Keil, 1987). Due to limited exposure, children have imperfect and partial knowledge which affects their concept representations. With added experience, their featural concept representations become increasingly accurate and differentiated. For example, young children tend to over-emphasize perceivable surface features of concepts (e.g., they describe the concept *uncle* as a person who is `a friend of the family` and `frequently brings presents`; Keil 1987). Over time, children learn the true defining features of the concept (i.e, that an *uncle* is the `brother of one parent`).

Another example of flexibility in meaning representation concerns diachronic change of word meaning. Language is a dynamic system that constantly adapts to changes in the cultural, economic or technical environment of its speakers (Traugott and Dasher, 2001). New words are established, for example in the context of technological innovation (e.g., the verb *to google*), meanings of words are extended (e.g., for about ten years the noun *tweet* has been used to refer to a short digital message) or restricted (e.g., the word *meat* was originally used to refer to 'food' in general).

In this chapter we investigate meaning development of individual concepts. We develop SCAN, a dynamic Bayesian model of Sense ChANge, which captures concept meaning as a set of senses whose changing nature is tracked over time. We model time as a sequence of discrete contiguous intervals and infer a meaning representation for each interval. Our model captures temporal variation *within* senses as well as change *across* senses, in their relative importance. We explicitly model the smooth and gradual nature of meaning change by enforcing that temporally adjacent meaning representations are co-dependent.

We apply SCAN to two phenomena of dynamic development of conceptual representations.[1] First, we investigate the change in concept representation in young children over time. We expose SCAN to input stimuli extracted from transcribed speech directed to children between one and five years in age, and monitor the development of meaning representations with increasing age. We show that the learnt temporal representations capture how premature child-like conceptual meanings develop towards more accurate and nuanced representations. To the best of our knowledge, we present the first computational study of meaning development in infants from naturalistic input. We investigate thematically broad featural patterns for a variety of natural concepts.

In addition, we use SCAN to study diachronic change of natural language (McMahon, 1994). Specifically, we monitor semantic change of individual words over centuries. We show that our model is able to detect changes across word senses like the emergence of a new sense (e.g., the word *mouse* in the mid-20th century acquired a new

[1] A note on terminology: In both applications, our model will be presented with textual input in the form of target terms in local context. In our child concept acquisition study, we will refer to target terms as *concepts* and to their meaning representation as sets of `feature types` (e.g., a *mouse* has feature types such as `appearance` or `behavior`). In this evaluation, *concepts* are nouns referring to living or non-living things, on the basic category level. In the diachronic language change study, we will refer to target terms as *words* and their representation as sets of `senses` (e.g., the word *mouse* has a sense relating to `animals` and a sense relating to `technical device`). We will use *words* comprising an a priori unrestricted set of nouns and verbs in our experiments. This distinction in terminology reflects the conventions in the NLP and cognitive literature, respectively, and makes it easier to discuss our work in the respective contexts.

sense relating to a computer device). Moreover, it infers subtle changes within a single sense (e.g., in the 1970s the words {cable, ball, mouse pad} were typical for the `computer device` sense, whereas nowadays the terms {optical, laser, usb} are more typical). We expose our model to a large text corpus of historical documents, covering more than three centuries, and show that it performs competitively on a range of meaning change detection tasks whilst inducing discernible word senses and capturing their development over time.

The remainder of this chapter is structured as follows. We motivate the two applications of SCAN and position it in the context of prior work in Section 6.1. Section 6.2 introduces our model formally and presents an approximate algorithm for parameter estimation. Section 6.3 presents our experiments on concept meaning development in children, and in Section 6.4 we evaluate SCAN on a variety of tasks relating to diachronic word meaning change. Section 6.5 summarizes our findings.

6.1 The Dynamic Nature of Meaning

We present prior experimental work on the acquisition and development of concept representations, and their change over time in Section 6.1.1. Section 6.1.2 reviews previous work on capturing diachronic word meaning change. Both sections include a review of previously proposed *computational* models, and position our own work in this context.

6.1.1 Acquisition and Development of Concept Representations

One of the most fundamental and challenging problems a young child is confronted with is to associate all and only relevant features with objects and concepts in her environment: which properties define an object to be a *ball*? Is a round candle a ball? What makes an animal a *dog*? Should it be alive? Does it have to possess a tail? Does it have to live next door? Is it called Bello? In this chapter we investigate the process with which featural representations of concepts develop in infants over time.

We review evidence for the dynamic nature of cognitive featural representations with a particular focus on developmental patterns during concept acquisition in infants. We conclude with an overview of computational models of human feature learning (Section 6.1.1.1).

The Dynamic Nature of Features The way humans acquire and use features suggests that cognitive featural representations are flexible and susceptible to change. First and foremost, children develop increasingly accurate concept representations over time. Feature learning and refinement is not unique to children, but similarly occurs in adults when they acquire new skills, e.g., in the process of specialist training, such as learning to distinguish a healthy from a broken bone in X-ray scans (Schyns and Rodet, 1997; Schyns et al., 1998; Norman et al., 1992). In this sense adults can be viewed as 'experts' and children as 'novices' in the context of learning to categorize and conceptually represent natural objects. In addition to individuals, societies create new meanings and shift the meaning of the linguistic concepts they use in communication in order to accommodate changes in their environment and communicative needs. We turn to this phenomenon in detail in Section 6.1.2.

Category representations do not only change diachronically, but also depend on the local situational context, e.g., the set of additional concepts present in a scene. Tversky (1977) showed that features used in similarity ratings change based on the set of concepts at hand: asked which of {Sweden, Hungary, Poland} is most similar to Austria, participants respond with Sweden (based on neutrality in the Cold War). When the same question is asked about {Sweden, Hungary, Norway} participants choose Hungary (based on geographical proximity). Situational context also influences the relevance of different features (e.g., the central features of an *apple* in a still life painting class likely differs from its central features in a lunch break context). This phenomenon has received attention under the term of *cross-categorization* and has been investigated both from a behavioral (Barsalou, 1987) and computational (Shafto et al., 2011) perspective. Finally, established categories have an influence on the featural representation of their members, as discussed in detail in Chapter 5.

The majority of behavioral studies (and computational models, see Section 6.1.1.1) on the development of representations in human learning focuses on *perceptual* features. These studies either investigate naturalistic adult learning processes such as learning to detect injuries in X-ray scans (Schyns and Rodet, 1997; Norman et al., 1992) or they are based on controlled laboratory experiments where participants learn discriminative features for a set of abstract shapes (such as 2-d or 3-d computer rendered images with a controlled feature space). One notable exception is Austerweil and Griffiths (2011), who experimentally investigate the acquisition of *conceptual* features in adults. They expose participants to facts about “Martian animals” found on a meteorite and investigate the development of the conceptual distinctions participants infer.

Conceptual Development in Children Common experience with conceptual mistakes young children make suggests that their concepts and categories blatantly diverge from those established among adults in a society. The differences in child and adult representations as well as the reliable convergence of child representations towards adult representations have received much attention in cognitive science and psychological research in the past (see e.g., (Meadows, 2006, Chapter 2.12) for a recent overview). Although initial research suggested that children are unable to create consistent categories given a set of ‘sortable’ objects (such as blocks of different size, shape or color, Inhelder and Piaget 1964), more recent results indicate that children do possess coherent concepts. These representations, however, differ from adult representations with regard to the salience given to different aspects of concept meaning, which result from children’s limited general knowledge (Mervis, 1987).

A range of behavioral studies found patterns of inaccuracies in child-like concept representations. On the basic category level, children (a) form categories that are broader than adults’ (e.g., round objects are *balls*, Chapman et al. 1986); (b) form categories that are narrower than adults’ (e.g., only the blue toy car is a *car*); or (c) form categories that partially overlap with adult categories (e.g., *cars* include *trains* but exclude *dune buggies*, Mervis 1987). Other work has investigated the development of featural category representations and their structure. Goldstone et al. (2001) found that children conflate features that adults distinguish (e.g., shade and size).

Keil (1987) investigates the structure of featural representations for different kinds of categories (animate, artifact and nominal), and their development. For socially constructed (nominal) concepts like *uncle* or *tax* he finds that the set of features shifts from characteristic to defining.[2] Children of different ages were presented with two kinds of definitions of concepts such as *uncle*: the definition either (a) contained many typical (`brings presents for Christmas`) but no necessary (`is brother of mother`

[2] Characteristic features are properties which are highly associated with a concept, but not necessary (e.g., although an *island* typically `has sandy beaches`, it remains an *island* even if this property is absent). Defining features are essential and their lack would change the identity of the concept (e.g., a piece of land that is not `surrounded by water` cannot be an *island*).

`or father`) features; or (b) contained necessary, and in addition very untypical (`is 2 years old`) features. Results show that children with increasing age gradually shift from preferring definition (a) to preferring definition (b). Similar patterns emerged for categories of animates (i.e., living things), and artifacts through 'object transformation' studies. Children of various ages were presented with scenarios in which perceptual features of animate concepts were drastically changed (e.g., "[...] one takes a raccoon, fluffs up his tail, sews a smelly sack inside, and even trains it to secrete the contents when alarmed [...]" (Keil, 1987, p. 187). While younger children judge the animal to change categories as a result of this transformation, older children and adults agree that despite the drastic changes the animal is still a raccoon. (Gelman, 1988) confirms the observed behavioral patterns for additional concepts and categories.

Why do the patterns of inaccuracies discussed above emerge, and how do child-like representations eventually approach adult-like representations? Various explanations and theories are offered in the literature. One salient argument concerns the lack of general knowledge, or theories, that underlie child category representation (Murphy and Medin, 1985). Young children do not know that the *essence* of animates (such as raccoons) is captured in their DNA and that changing the raccoon's perceptual features will not change its fundamental property of being a raccoon. Similarly young children are not aware of family relations that define the term *uncle* (Keil, 1987). They thus rely on *surface* features which are prevalent and perceptible. Additionally, due to a lack of experience, false beliefs or false features might temporarily influence the child's categorization (e.g. a leopard, which looks similar to a cat might say "meow", Mervis 1987). Finally, children may weigh features differently due to the limited set of situations they have been exposed to (Mervis, 1987): although they know all the properties of the concept *island*, the features {water, beach, holiday} might be salient in the child's representation so that she temporarily classifies all places that resemble a beach resort as *islands* (Keil, 1989, Chapter 4). Gelman (1988) shows an increased influence of domain-specific knowledge on category-specific inferences in school children when compared to pre-schoolers.

The development from child-like conceptual representations towards adult-like representations has been shown to be individual. The characteristics of the child's environment, i.e., the salience of different concepts in the surroundings influences the speed and order of learning (Neisser, 1987). In addition, the feedback received through interactions with adults has been shown to have an influence on the learning process (e.g., through explicit illustration of important properties of objects, or acknowledging relevant properties that the child discovered herself, Mervis 1987). We will return to individual developmental difference in feature learning in the analysis of our results.

The influence of language on the conceptual development in children has been the subject of debate and is difficult to pin down exactly (see (Goswami, 2014, Chapter 3) for a discussion). Does increasing linguistic competence *change* mental concept representations, or merely facilitate their communication? Evidence suggests that language supports the acquisition of imperceptible, knowledge-based features, which allow children to learn conceptualizations that go beyond perceptual representation (see also our discussion in Section 2.1). Experiments with 2-year old children showed that linguistic labels (e.g., 'bird') improve feature prediction accuracy when perceptual cues are not informative (e.g., for atypical category members, Gelman and Coley 1990). In the absence of a label 2-year olds tend to predict features based on perceptual similarity (e.g., bird features are predicted for a dinosaur perceptually similar to a bird rather than for atypical birds like pelicans). Similar studies showed that 3-4 year olds do not require the label and are able to use structural knowledge immediately, which prevents them from relying on misleading perceptual cues (Gelman and Markman, 1986, 1987).

Our experiments (Section 6.3) investigate the extent to which the developmental patterns discovered in behavioral studies are captured by our computational model of dynamic sense change. We investigate this question from a *modeling*, and now position our work in the context of previously proposed models of human feature learning.

6.1.1.1 Computational Models of Human Feature Learning

A variety of computational accounts for human feature learning have been proposed ranging methodologically from neural networks to Bayesian methods, covering both the acquisition of perceptual (surface) features and conceptual (underlying relational, or knowledge-based) features. We begin our overview with low-level neural network models of perceptual feature learning, and proceed to describing higher-level Bayesian methods and computational models of conceptual feature learning.

Kruschke (1992) model perceptual feature acquisition using a neural network which learns to weigh features based on their utility for concept categorization. The model is based on an exemplar categorization model and assumes that features emerge as a result of concept categorization. The original model was constructed for perceptual stimuli represented as points in a feature space and has subsequently been extended to account for binary featural representations of stimuli indicating the presence or absence of particular features (Lee and Navarro, 2002).

Modeling feature learning purely based on categorization performance does not capture the fact that statistical co-occurrence patterns of features across objects also play an important role in feature acquisition. Goldstone (2003) and Goldstone et al. (2008) develop a neural network model for perceptual feature learning. In addition to the conceptual bias (features should explain category membership of concepts) their model includes a distributional bias: features that are spatially close in the stimulus (e.g., adjacent pixels) should receive similar feature values.

Both models discussed above learn features by re-weighting an existing inventory of features. Fahlman and Lebiere (1990) propose a method for incrementally changing the structure of a neural network such that additional nodes can be added to account for increasing complexity in the input data. Love et al. (2004) use this method to model incremental category learning.

A similar idea of adaptively increasing model complexity with the complexity of the structure in the input has been put forward in the form of non-parametric Bayesian models. A series of rational (aka ideal learner) Bayesian models have been proposed which infer perceptual features from raw pixel input, and account for a variety of cognitive phenomena in human concept learning such as the influence of categorization on feature creation and incremental learning (Austerweil and Griffiths, 2009, 2011, 2013). These models formalize feature learning as non-parametric Bayesian inference of the simplest set of features that explains the input stimuli. They incorporate a simplicity-encouraging prior over features in form of the Indian Buffet Process (IBP).

The models discussed so far largely focus on the acquisition of perceptual features for a set of visual concepts. The stimuli themselves tend to be limited in their feature complexity (e.g., 2-dimensional line drawings with black or white pixels, or 3-dimensional computer rendered gray scale images). They do not resemble the visual input a child encounters, namely cluttered scenes with a potentially unbounded number of objects of varying complexity.

Beyond computational models of visual feature learning, a limited number of models for learning conceptual features have been proposed as well. Austerweil and Griffiths (2011) apply their IBP-based models discussed above to adult acquisition of conceptual categories: They present participants with descriptions of instances of novel species present in a set of "Martian fossils", and investigate the process with which participants identify discriminating features for the species. While this work is conceptually most similar to our goal of modeling conceptual feature learning, it focuses on feature development in adults, who are equipped with substantial prior experience with concept representations and categorization. The inherent limitations of laboratory experiments restrict the scope of their setup to a small number of concepts and features. It is not clear how the results extend to a more naturalistic setting of learning on a larger scale with potentially unlimited features.

Zeigenfuse and Lee (2010) present a Bayesian model of feature learning from a large set of human-produced feature norms and similarity ratings for a set of domain-specific concepts (e.g., animals). Their model extracts a subset of the input features that explain the feature similarity ratings well. They assume an underlying feature learning process that optimizes the distance of concepts in a representational space, where each concept is represented by a weighted vector of features. Other work (Perfors et al., 2005) investigated the development of features and structured domain knowledge and their interaction in the context of concept acquisition in children. They develop a Bayesian model which infers adequate domain-specific knowledge structure (e.g., hierarchical vs flat) from a set of binary-featured object-feature matrices for the domains of *food* and *animal*. Both models infer features from an already highly constrained feature set (based on human produced features). We model the acquisition from noisy input data in the form of transcribed child-directed speech.

Finally, our model is related to computational models of word learning. Young children learn new nouns with a rapid pace, and it has been shown that knowledge about correlating properties facilitates this process (Jones et al., 1991; Landau et al., 1998). Various models have been proposed that explore the interplay of word learning and category learning, comprising both connectionist (Colunga and Smith, 2005; Colunga and Sims, 2011) as well as probabilistic approaches (Yu, 2005). See Section 2.2 for a thorough review of computational models of word learning and their relation to concept and category acquisition.

Experiment 5 in Section 6.3 explores the development of feature representations of *concepts* in the form of thematically coherent clusters of words which change over time in their nature and relevance, from child-directed language. We advance previous work in three ways: First, we are interested in the development of *conceptual* features of living and non-living things. Secondly, we investigate the process of how *young children* acquire features for basic level concepts. Finally, we model the acquisition process in a more *natural setting* that previous work, based on the statistical regularities in natural language input available to the child. To the best of our knowledge we present the first large-scale computational study of conceptual development in children.

6.1.2 Diachronic Meaning Change of Words

We now turn to meaning development on a larger scale. We will use SCAN to track and analyze how language changes over decades or centuries. We will investigate diachronic low-level semantic change of the meaning of individual words.

Language is a dynamic system, constantly evolving and adapting to the needs of its users and their environment (Aitchison, 2001). Words in all languages naturally exhibit a range of senses whose distribution or prevalence varies according to the genre and register of the discourse as well as its historical context. As an example, consider the word *cute* which according to the Oxford English Dictionary (OED, Stevenson 2010) first appeared in the early 18th century and originally meant `clever` or `keen-witted`. By the late 19th century *cute* was used in the same sense as *cunning*. Today it mostly refers to objects or people perceived as `attractive`, `pretty` or `sweet`. Another example is the word *mouse* which initially was only used in the `rodent` sense. The OED dates the `computer pointing device` sense of *mouse* to 1965. The latter sense has become particularly dominant in recent decades due to the ever-increasing use of computer technology.

The arrival of large-scale collections of historic texts (Davies, 2010) and online libraries such as the Internet Archive and Google Books have greatly facilitated computational investigations of language change. The ability to automatically detect how the meaning of words evolves over time is potentially of significant value to lexicographic and linguistic research but also to real world applications. Time-specific knowledge would presumably render word meaning representations more accurate, and benefit downstream tasks where semantic information is crucial. Examples include information retrieval and question answering, where time-related information could increase the precision of query disambiguation and document retrieval (e.g., by returning documents with newly created senses or filtering out documents with obsolete senses).

6.1.2.1 Computational Models of Diachronic Meaning Change

Most work on diachronic language change has focused on detecting whether and to what extent a word's meaning changed (e.g., between two epochs) without identifying word senses and how these vary over time. A variety of methods have been applied to the task ranging from the use of statistical tests in order to detect significant changes in the distribution of terms from two time periods (Popescu and Strapparava, 2013; Cook and Stevenson, 2010), to training distributional similarity models on time slices (Gulordava and Baroni, 2011; Sagi et al., 2009), and neural language models (Kim et al., 2014; Kulkarni et al., 2015). Other work (Mihalcea and Nastase, 2012) takes a supervised learning approach and predicts the time period to which a word belongs given its surrounding context.

Bayesian models have been previously developed for various tasks in lexical semantics (Brody and Lapata, 2009; Ó Séaghdha, 2010; Ritter et al., 2010) and word meaning change detection is no exception. Using techniques from non-parametric topic modeling, Lau et al. (2012) induce word senses (aka topics) for a given target word over two time periods. Novel senses are then are detected based on the discrepancy between sense distributions in the two periods. Follow-up work (Cook et al., 2014; Lau et al., 2014) further explores methods for how to best measure this sense discrepancy. Rather than inferring word senses, Wijaya and Yeniterzi (2011) use a Topics-over-Time model and k-means clustering to identify the periods during which selected words move from one topic to another.

A non-Bayesian approach is put forward in Mitra et al. (2014, 2015) who adopt a graph-based framework for representing word meaning (see Tahmasebi et al. (2011) for a similar earlier proposal). In this model words correspond to nodes in a semantic network and edges are drawn between words sharing contextual features (extracted from a dependency parser). A graph is constructed for each time interval, and nodes are clustered into senses with Chinese Whispers (Biemann, 2006), a randomized graph clustering algorithm. By comparing the induced senses for each time slice and observing inter-cluster differences, their method can detect whether senses emerge or disappear.

Our work draws ideas from dynamic topic modeling (Blei and Lafferty, 2006b) where the evolution of topics is modeled via (smooth) changes in their associated distributions over the vocabulary. Although the dynamic component of our model is closely related to previous work in this area (Mimno et al., 2008), our model is specifically constructed for capturing sense rather than topic change. Our approach is conceptually similar to Lau et al. (2012). We also learn a joint sense representation for multiple time slices. However, in our case the number of time slices in not restricted to two and we explicitly model temporal dynamics. Like Mitra et al. (2014, 2015), we model how senses change over time. In our model, temporal representations are not *independent*, but influenced by their temporal neighbors, encouraging smooth change over time. We therefore induce a *global* and consistent set of temporal representations for each word. Our model is knowledge-lean (it does not make use of a parser) and language independent (all that is needed is a time-stamped corpus and tools for basic pre-processing). Contrary to Mitra et al. (2014, 2015), we do not treat the tasks of inferring a semantic representation for words and their senses as two separate processes.

Our evaluation in Section 6.4 reveals that SCAN (a) induces temporal representations which reflect word senses and their development over time, (b) is able to detect meaning change between two time periods, and (c) is expressive enough to obtain useful features for identifying the time interval in which a piece of text was written. Overall, our results indicate that an explicit model of temporal dynamics is advantageous for tracking meaning change. Comparisons across evaluations and against a variety of related systems show that despite not being designed with any particular task in mind, our model performs competitively across the board.

6.2 A Dynamic Bayesian Model of Semantic Change

In this section we introduce SCAN, our dynamic Bayesian model of Semantic ChANge. SCAN induces a globally coherent representation of meaning development of individual words over time[3], comprising a set of time-specific word meaning representations. We start by explaining the intuitions and assumptions underlying our model, and continue with a technical model description, before we describe an approximate learning algorithm.

[3] Throughout the model description we use the term *word* to refer to target terms whose meaning change we aim to model. This refers either to linguistic words in the diachronic language change evaluation, or to mental concepts in the concept development study. Similarly, we use the term *sense* throughout the model description, which will correspond either to word senses (of linguistic words) or to feature types (of mental concepts).

Intuition We create a SCAN model for an individual target word c which captures the development of its meaning over time. The input to the model is a corpus of short text snippets (or documents), each consisting of a mention of the target word c and its local context $\mathbf{w}$ as a symmetric context window of $\pm n$ words. Each snippet is annotated with its corresponding time stamp. This corresponds to the age of the addressed child for the cognitive development experiments, and the year of the document's origin for the diachronic meaning change experiments. Example input documents are displayed in Tables 6.1 (page 183) and 6.4 (page 197).

Given such a set of input documents, how do we model word meaning and its temporal dynamics? We represent the *meaning of a word* as a set of senses. Each sense captures an internally coherent aspect of its meaning, and is characterized through a set of words that are associated with that sense. Senses are further distinguished in terms of their prevalence since not all meanings are equally common for each word at all times. We assume that each input text snippet refers to exactly one sense. We formalize *temporal dynamics* assuming a discrete set of contiguous time intervals. Given a target word whose meaning development is to be tracked, our model infers a meaning representation for each time interval. We introduce dependencies between temporally adjacent time-specific meaning representations so as to explicitly capture the gradual nature of meaning change with respect to both sense prevalences and sense-characterizing words.

The output of a SCAN model is a globally coherent set of time interval-specific word representations comprising the prevalence and content of word senses over time. Individual representations are inferred jointly, capturing meaning change as a smooth process, and inducing a globally meaningful and coherent picture of word meaning.

Model Description We now describe SCAN more formally. The generative story of our model is displayed in Figure 6.1a and its plate diagram representation can be found in Figure 6.1b.

A SCAN model is parameterized with regard to the number of senses $k \in [1...K]$ of the target word c, and the length of time intervals ΔT which might be finely or coarsely defined (e.g., spanning a month, a year, or a decade). We conflate all inputs originating from the same time interval $t \in [1...T]$ and infer a temporal representation of the target word per interval. We use $v \in [1...V]$ as an index over the vocabulary. A temporal

(a) Generative story of SCAN.

```
Draw $\kappa^{\phi} \sim Gamma(a,b)$
for time interval $t = 1..T$ do
    Draw sense distribution    $\phi^t|\phi^{-t},\kappa^{\phi} \sim \mathcal{N}(\frac{1}{2}(\phi^{t-1}+\phi^{t+1}),\kappa^{\phi})$
    for sense $k = 1..K$ do
        Draw word distribution    $\psi^{t,k}|\psi^{-t},\kappa^{\psi} \sim \mathcal{N}(\frac{1}{2}(\psi^{t-1,k}+\psi^{t+1,k}),\kappa^{\psi})$
    for document $d = 1..D$ do
        Draw sense $z^d \sim Mult(\phi^t)$
        for context position $i = 1..I$ do
            Draw word $w^{d,i} \sim Mult(\psi^{t,z^d})$
```

(b) Plate diagram of SCAN.

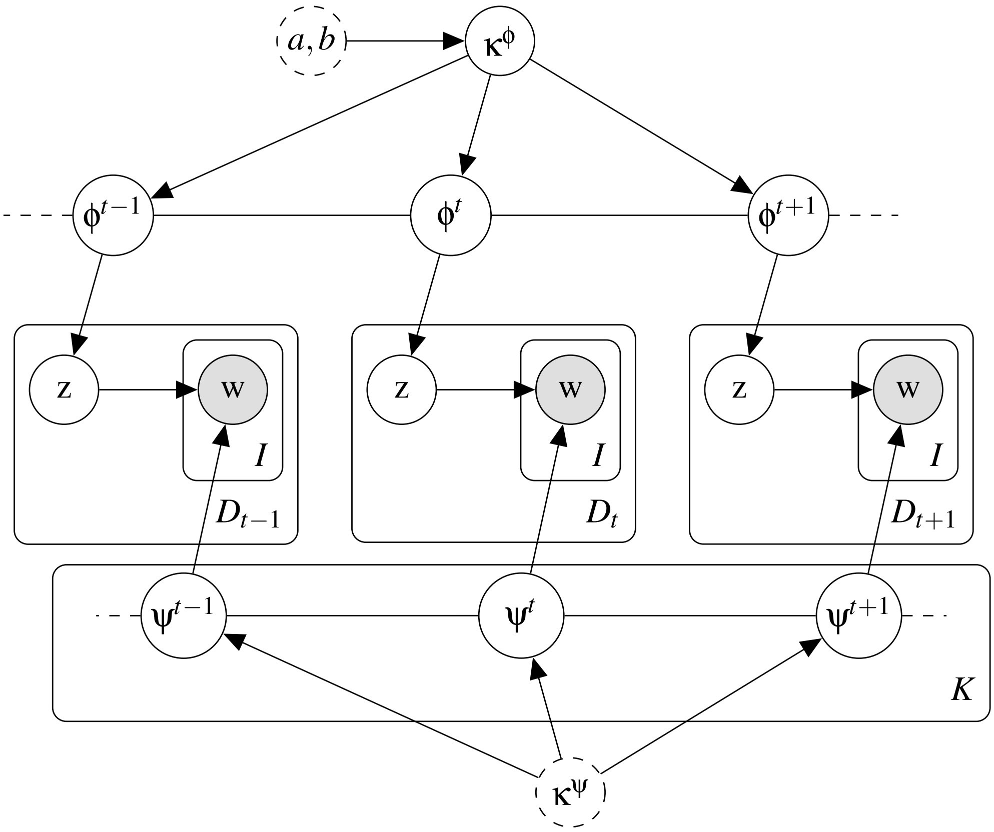


Figure 6.1: Top (a): The generative story of SCAN. Observations (w) and latent labels (z) are drawn from Multinomial distributions ($Mult$). Parameters for the multinomial distributions are drawn from logistic normal distributions ($\mathcal{N}$). Bottom (b): The plate diagram representation of SCAN for three time steps $\{t-1,t,t+1\}$. Constant parameters are shown as dashed nodes, latent variables as clear nodes, and observed variables as gray nodes.

meaning representation of word c at each time t comprises:

- the relative prevalence of senses at that time, as a K-dimensional multinomial distribution over senses ϕ^t, and
- a representation of each sense k at time t as a V-dimensional multinomial distribution over the vocabulary $\psi^{t,k}$.

Intuitively, the temporal meaning representations are not independent of each other, but develop dynamically over time, each depending on their temporal neighbors. We encode this intuition into the prior distributions, embedding them into a time series model and 'tieing together' the values of each individual multinomial parameter ϕ^t_k and $\psi^{t,k}_w$ with its temporal neighbors at times $t-1$ and $t+1$. We technically describe this prior in Section 6.2.1.

The generative story of SCAN (Figure 6.1a) proceeds as follows. Each SCAN model captures the meaning change of one given target word c. First, we draw a precision parameter (κ^ϕ) from its prior *Gamma* distribution, which regulates the global extent of sense prevalence change over time for target word c. For each time interval t we draw parameters of a Multinomial distribution over senses from the logistic normal prior (ϕ^t, capturing each sense's prevalence). For each time interval t and each sense k we draw a set of Multinomial parameters over the vocabulary, from a separate logistic normal prior ($\psi^{t,k}$, capturing each sense's content). Next, we generate time-specific text snippets (or documents). For each snippet d, we first observe its time stamp t, and generate a sense from the time-specific Multinomial sense distribution ϕ^t. Finally, we draw a fixed number of context words independently from the sense-specific Multinomial distribution over words for time t, $\psi^{t,k}$.

6.2.1 The Time Series Prior

We define the prior of the SCAN model in a way that allows us to 'tie' Multinomial parameterizations across neighboring time steps, i.e., to capture the smooth nature of meaning change. The Dirichlet distribution is the most common choice of a prior distribution in a model with Multinomial data-generating distributions due to mathematical convenience. However, it is limited in the dependencies it can encode between parameters. Instead we draw our multinomial parameters from the logistic normal distribution (Aitchison, 1982; Blei and Lafferty, 2006a), and embed these distributions in a time series model. We first describe the parameter-generating process, and then explain how we embed the logistic normal prior distributions into a time series model.

The Logistic Normal Distribution. A draw from the logistic normal distribution consists of:

(1) a draw of an n-dimensional random vector from the multivariate normal distribution parameterized by mean vector μ and variance-covariance matrix Σ, $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, and

(2) a mapping of the drawn parameters to the simplex through the logistic transformation $\theta_i = exp(x_i)/\sum_{i'} exp(x_{i'})$.

The logistic transformation ensures that θ is a valid set of multinomial parameters (i.e., that $0 \leq \theta_i \leq 1\ \forall \theta_i$ and $\sum_i \theta_i = 1$). The parameterization of the multivariate normal distribution (through mean vector μ and variance-covariance matrix Σ) allows to encode structured prior knowledge, such as correlation of parameters (Blei and Lafferty, 2006a). The distribution can also be straightforwardly integrated into time series models (Blei and Lafferty, 2006b; Mimno et al., 2008), which is our goal here.

We follow the two-stage procedure explained above, and draw for each time t a K-dimensional random vector from the logistic normal prior over the sense prevalence distributions:

$$\beta \sim \mathcal{N}(\mu^{\phi}, \Sigma^{\phi}), \tag{6.1}$$

and deterministically compute the multinomial parameters ϕ^t through the logistic transformation:

$$\phi^t_k = \frac{exp(\beta_k)}{\sum_{k'} exp(\beta_{k'})}. \tag{6.2}$$

Equivalently, for each time t we draw parameters independently for each sense-specific multinomial distribution over the vocabulary:

$$\begin{aligned} \gamma &\sim \mathcal{N}(\mu^{\psi}, \Sigma^{\psi}) \\ \psi^{t,k}_v &= \frac{exp(\gamma_w)}{\sum_{v'} exp(\gamma_{v'})}. \end{aligned} \tag{6.3}$$

We assume that given a time t all individual sense prevalences ϕ^t_k, as well as all sense-specific word probabilities $\psi^{t,k}_v$ are independent of each other. Hence we define Σ^{ϕ} and Σ^{ψ} as diagonal matrices where the value along the diagonal will correspond to κ^{ϕ} and κ^{ψ}, respectively (see below). We assume zero mean vectors $\mu^{\phi} = \mu^{\psi} = 0$.

Modeling temporal dynamics. We embed the logistic normal distributions in a dynamic model which captures the development of meaning through temporally local dependencies between multinomial parameters, and encourages smooth change. We model the dynamics of meaning development over time in SCAN through intrinsic Gaussian Markov Random Fields (iGMRFs; Rue and Held 2005). This section provides a brief reminder of iGMRFs in the context of our model. Please refer to Section 3.2.3 for a more general introduction and motivation for their use as priors in Bayesian models. An exhaustive introduction to GMRFs can be found in Rue and Held (2005) (see also Vivalt (2014) for an accessible overview; for an application of iGMRFs to topic models, see Mimno et al. (2008)).

Let $\phi = \{\phi^1 ... \phi^T\}$ denote a T-dimensional random vector, where each ϕ^t might for example correspond to the probability of a sense at time t.[4] We define a prior which encourages smooth change of parameters at neighboring times, in terms of a first-order random walk on the line (graphically depicted as the chains of ϕ and ψ in Figure 6.1b). Specifically, we define this prior as an intrinsic Gaussian Markov Random Field, which allows us to model the *change* of adjacent parameters as drawn from a normal distribution,[5] e.g.:

$$\Delta\phi^t \sim \mathcal{N}\Big(0, \frac{1}{\kappa}\Big), \tag{6.4}$$

where we assume zero mean and κ is the precision, i.e., the inverse of the variance. iGMRFs are defined with respect to the parameter chains ϕ and ψ^k, respectively (Figure 6.1b); it is sparsely connected with only first-order dependencies which allows for efficient inference. A second feature, which makes iGMRFs popular as priors in Bayesian modeling, is the fact that they can be defined purely in terms of the local changes between dependent (i.e., adjacent) variables, without the need to specify an overall mean of the model. The full conditionals explicitly capture these intuitions (cf. Section 3.2.3 for technical details):

$$\phi^t | \phi^{-t}, \kappa \sim \mathcal{N}\Big(\frac{1}{2}(\phi_{t-1} + \phi_{t+1}), \frac{1}{2\kappa}\Big), \tag{6.5}$$

for $1 < t < T-1$, where ϕ^{-t} is the vector ϕ except element ϕ^t and κ is a precision parameter. The value of parameter ϕ^t is distributed normally, centered around the mean of the values of its neighbors, without reference to a global mean. The precision

[4] This is the simplest case, modeling the development of one sense. In our model each ϕ^t is a K-dimensional vector, specifying a probability distribution over K senses.

[5] In what follows we assume a zero mean and leave the diagonal variance-covariance matrix Σ implicit.

parameter κ controls the extent of variation: how tightly coupled are the neighboring parameters? Or, in our case: how tightly coupled are temporally adjacent meaning representations of a word c?

Hyperparameter Sampling. The hyperparameters κ^{ϕ} and κ^{ψ} control the degree to which prevailing senses and sense-specific word distributions are allowed to vary over time. We estimate the value of κ^{ϕ} during inference, which allows us to model the extent of temporal change in prevalence of senses individually for each target word. We draw κ^{ϕ} from a conjugate Gamma prior $\kappa^{\phi} \sim Gamma(a,b)$ with shape parameter a and rate parameter b. We do not infer the sense-word precision parameter κ^{ψ}. Instead, we fix it at a high value, triggering little variation of word distributions within senses. This leads to individual senses being thematically consistent over time, making sure that we track the development of senses that refer to the same aspect of a target word's meaning throughout.

In summary, given a corpus of D documents, we wish to infer the following latent variables:

(1) sense assignments to documents $\{z\}^{D}$,

(2) time-specific sense distributions $\{\phi\}^{T}$,

(3) time- and sense-specific word distributions $\{\psi\}^{T\times K}$, and

(4) the sense precision parameter κ^{ϕ}.

The full posterior distribution over latent variables given the data w, parameters a, b, κ^{ψ}, and the choices of distributions described above factorizes as,

$$
\begin{aligned}
&P(z,\phi,\psi,\kappa^{\phi}|w,\kappa^{\psi},a,b)\\
&\quad = P(\kappa^{\phi}|a,b)P(\phi|\kappa^{\phi})P(\psi|\kappa^{\psi})P(z|\phi)P(w|z,\psi)\\
&\quad \propto Ga(\kappa^{\phi};a,b)\prod_{t}\Big[\mathcal{N}(\phi^{t}|\kappa^{\phi})\prod_{k}\big[\mathcal{N}(\psi^{t,k}|\kappa^{\psi})\big]\times\\
&\qquad\qquad\qquad\qquad\qquad \prod_{d}\big[Mul(z|\phi^{t})\prod_{i}Mul(w^{i}|\psi^{z,t})\big]\Big]\\
&\quad = Ga(\kappa^{\phi};a,b)\prod_{t}\left[\mathcal{N}(\phi^{t}|\kappa^{\phi})\prod_{k}\left[\mathcal{N}(\psi^{t,k}|\kappa^{\psi})\right]\prod_{d}\left[\phi^{t}_{z}\prod_{i}\psi^{z,t}_{w^{i}}\right]\right],
\end{aligned}
\tag{6.6}
$$

where we use Ga to refer to the Gamma distribution, Mul to refer to the multinomial distribution, and $\mathcal{N}$ to refer to the Logistic Normal distribution obeying the structural

Algorithm 6 The Gibbs sampling algorithm for the SCAN model.

```
Input: model with randomly initialized parameters.
Output: posterior estimate of $\mathbf{z}, \phi, \psi, \kappa^\phi$
repeat
    for document $d$ do                                    ▷ Sample sense assignments $\mathbf{z}$
        $z^d \sim p(z^d = k|w^-, z^-, \phi, \psi) = \phi^{k,t} * \prod_f \left(\psi_f^{k,t}\right)^{N_f^d}$
    for time $t$ do                                        ▷ Sample sense parameters $\phi$
        $\beta_k \sim p(\beta|\mathbf{z}) \propto \prod_k \left(\frac{\exp(\beta_t^k)}{\sum_{k'} \exp(\beta_{k'})}\right)^{N_t^k} \mathcal{N}\left(\beta_k;, \kappa^\phi\right)$
        $\phi^t = \text{logistic-transform}(\beta)$
    for time $t$ do                                        ▷ Sample word parameters $\psi$
        for sense $k$ do
            $\gamma_w \sim p(\gamma|\mathbf{z}, \mathbf{w}) \propto \prod_f \left(\frac{\exp(\gamma_v^{k,t})}{\sum_{w'} \exp(\gamma_{w'}^{k,t})}\right)^{N_f^{k,t}} \mathcal{N}\left(\gamma_v^{k,t}; \kappa^\psi\right)$
            $\psi^{t,k} = \text{logistic-transform}(\gamma)$
                                                           ▷ Sample precision parameter $\kappa^\phi$
    $\kappa^\phi \sim p(\kappa^\phi|\phi)p(\kappa^\phi; a, b) = Ga\left(\frac{KT}{2} + a, \frac{1}{2}\sum_{t,s}\left(\phi_s^t - \frac{1}{2}(\phi_{t-1}^k + \phi_{t+1}^k))\right)^2 + b\right)$
until convergence
```

dependencies defined through the iGMRF prior.[6]

6.2.2 Batch Learning

We use a blocked Gibbs sampler for approximate inference, which repeatedly executes three steps which alternately resample (a) document-sense assignments, (b) multinomial parameters from the logistic normal prior, and (c) the sense precision parameter from a Gamma prior. The full sampling procedure is displayed in Algorithm 6.

The logistic normal prior is not conjugate to the multinomial distribution. This means that the form of the conditional posterior distributions over logistic normal parameters β and γ is unknown and cannot be sampled from straightforwardly. One way to alleviate the problem of sampling from an unknown or very complex distribution are auxiliary variable (or data augmentation) techniques. A set of auxiliary variables is drawn from a well-known distribution (the uniform distribution in our case), and the

[6] In order to keep notation to a minimum, we use $\mathcal{N}$ as a shorthand for the Logistic Normal distribution, comprising both the draw from the normal distribution and the logistic transformation.

parameters of interest are drawn conditioned on this set of auxiliary variables. This cascade is carefully set up such that the auxiliary variables do not change the underlying distributions of interest but only serve as helper variables to ease the computations.

We adapt the auxiliary variable sampler introduced in Mimno et al. (2008) to our model (see also Groenewald and Mokgatlhe (2005)), and describe the three components of our sampler in turn below.

Resampling document-sense assignments z. Our sampler first iterates over the input documents d (each consisting of a set of words $\mathbf{w}$), and resamples their sense assignments under the current model parameters $\{\phi\}^T$ and $\{\psi\}^{K\times T}$. Similarly to the approach taken in Gibbs sampling for Dirichlet-Multinomial models, each document label z^d is individually resampled given the current values of all other variables in the model.[7] We sample from its posterior distribution, combining the prior distribution over labels at time t with the likelihood of observing words w under this label at time t:

$$\begin{aligned} p(z^d|\mathbf{w},t,\phi,\psi) &\propto p\left(z^d|t\right)p\left(\mathbf{w}|t,z^d\right) \\ &= \phi^t_{z^d}\prod_i \psi^{t,z^d}_{w^i} \end{aligned} \tag{6.7}$$

Resampling multinomial parameters ϕ and ψ. Next, we resample parameters $\{\phi\}^T$ and $\{\psi\}^{K\times T}$ from the logistic normal prior, given the current sense assignments to the data. We use the auxiliary variable sampler proposed in Mimno et al. (2008). An illustration of the procedure is displayed in Figure 6.2.

Recall that Multinomial parameters ϕ (and ψ) are obtained by drawing vectors β (and γ) from the MVN and subsequently mapping them to the simplex. The parameter vectors β and γ are resampled independently and component-wise, and are subsequently re-normalized to yield the valid multinomial parameters ϕ and ψ. Intuitively, we will sample a new value for each individual β^t_k (and, equivalently, $\gamma^{t,k}_w$) from a bounded, weighted area (cf., Figure 6.2, center). The boundaries are determined by the current assignments of target sense k to documents from target time t (or, equivalently for $\gamma^{t,k}_w$, observations of target words w under sense k and time t). The weights of different values in the bounded area are determined by the iGMRF prior, triggering values to

[7] For a mathematical description of Gibbs sampling for Dirichlet-Multinomial models please refer to Section 3.3.2.3 (pages 47 ff.).

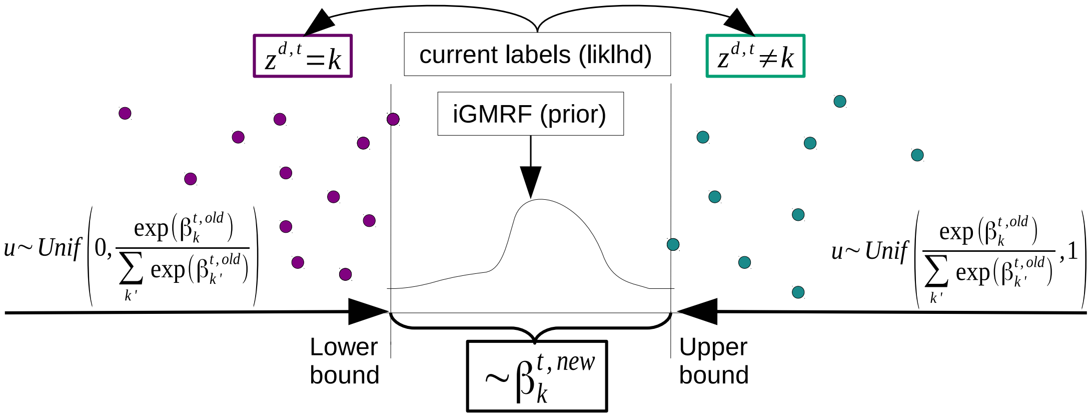

$z^{d,t}=k$
current labels (liklhd)
$z^{d,t}\neq k$
iGMRF (prior)
$u\sim Unif\left(0,\frac{\exp(\beta_k^{t,old})}{\sum_{k'}\exp(\beta_{k'}^{t,old})}\right)$
$u\sim Unif\left(\frac{\exp(\beta_k^{t,old})}{\sum_{k'}\exp(\beta_{k'}^{t,old})},1\right)$
Lower bound
$\sim\beta_k^{t,new}$
Upper bound

Figure 6.2: Schematic illustration of the process underlying the auxiliary variable sampling scheme for logistic-normal distributed parameters. We sample a new value for parameter $\beta_k^{t,new}$ based on its old value ($\beta_k^{t,old}$), the number of times sense k was assigned to any document from time slice t, and the iGMRF prior.

be similar to their temporally neighboring values, with the degree of similarity (or permitted flexibility) determined through the precision parameters κ^{ϕ} and κ^{Ψ}. We will describe the resampling procedure for one component β_k^t below, noting that the procedure for $\gamma_w^{t,k}$ follows the exact same reasoning.

We resample the prevalence parameter β_k^t, capturing the probability of sense k at time t, from a bounded weighted area. The boundaries of the weighted area are approximated using a set of auxiliary variables. Indeed, this approximation is identical to the approximation of the cumulative distribution function of a logistic distribution, which means that we can proceed as follows.[8] We draw an auxiliary variable for each document d from target time t. The value is drawn uniformly from an interval with boundaries depending on whether the document's sense assignment z_d corresponds to our target sense k (case 1) or not (case 2):

$$u_i \sim \begin{cases} \text{unif}\left(0, \frac{\exp(\beta_k)}{\sum_{k'}\exp(\beta_{k'})}\right) & \text{if } z_d = k \\ \text{unif}\left(\frac{\exp(\beta_k)}{\sum_{k'}\exp(\beta_{k'})}, 1\right) & \text{otherwise.} \end{cases} \tag{6.8}$$

The largest value drawn in the former case, $u_{max}^{z_d=k}$, will determine the lower bound of the region from which a new value for β_k^t will be drawn: the more documents at time t are already assigned sense k, the higher the lower bound is expected to be. This is illustrated through the swarm of violet sample points in Figure 6.2 (left). Conversely,

[8] Please consult Groenewald and Mokgatlhe (2005) and Mimno et al. (2008) for the mathematical details, which legitimate the approach, but are not necessary for an intuitive understanding of the method.

the lowest value drawn in the latter set of random variables (if $z_d \neq k$), $u_{min}^{z_d \neq k}$, will determine the upper bound of the region: the more documents are assigned senses other than the current target sense k, the lower the upper bound is expected to be (cf., Figure 6.2 (blue samples on the right)).[9]

Given these values, the new value for β_k^t is drawn from the area bounded by,

$$log \frac{\left(\sum_{k' \neq k} exp(\beta_k^t)\right) u_{max}^{z_d = k}}{1 - u_{max}^{z_d = k}} < \beta_k^t < log \frac{\left(\sum_{k' \neq k} exp(\beta_k^t)\right) u_{min}^{z_d \neq k}}{1 - u_{min}^{z_d \neq k}}. \tag{6.9}$$

The area within the boundaries is weighted with respect to the prior iGMRF as defined above (Equation 6.5). We thus draw the new value of β_k^t from a truncated normal distribution, with mean averaged over all dependent (i.e., adjacent) parameter values, and precision determined by κ^ϕ. The normal distribution is truncated at the bounds defined above. Finally, we deterministically update the parameter vector ϕ^t given the new value β_k^t using the logistic transformation.

To sum up, the resampled value of each individual β_k^t is determined by (a) the importance of sense k from the current sense assignments to documents from time t; and (b) the extent to which any new value agrees with the temporal coherence constraint imposed by the iGMRF prior.

Resampling the precision parameter κ^ϕ. Finally, we periodically resample the sense precision parameter κ^ϕ from its posterior distribution

$$\begin{aligned} p(\kappa^\phi|\phi, a, b) &\propto \prod_t \prod_k \mathcal{N}(\phi_k^t|\kappa^\phi) \times Ga(\kappa^\phi; a, b) \\ &\propto \prod_t \prod_k \mathcal{N}(\phi_k^t|\frac{1}{2}(\phi_k^{t-1} + \phi_k^{t+1}), \frac{1}{2\kappa^\phi}) \times \kappa_\phi{}^{a-1} \exp[-\kappa_\phi b], \end{aligned} \tag{6.10}$$

which is itself a Gamma distribution with parameters:

$$\kappa^\phi|\{\phi\}^t, a, b \sim Ga\left(\frac{KT}{2} + a, \frac{\sum_{t,k}\left(\phi_k^t - \frac{1}{2}(\phi_k^{t-1} + \phi_k^{t+1})\right)^2}{2} + b\right) \tag{6.11}$$

Intuitively, this represents the prior shifted by half the number of observations and half the sum of squared divergences from the mean.

This section presented the technical details of SCAN, and derived a blocked Gibbs sampler for approximate learning. In the remainder of the chapter, we apply our model to

[9] Mimno et al. (2008) explain various ways to make this procedure more efficient. We use these methods in our implementation, and refer the interested reader to their paper for additional details.

age (y-mm)	**utterance**		
1-01	day pajamas pajamas	*bed*	yawn stretch touch
1-05	bed brush brush	*bed*	brush teeth tooth
2-00	sleep tire book	*bed*	bed sit fall
2-06	snore under gentle	*bed*	sweet dream silly
3-00	wake early play	*bed*	awful mess upstairs
1-01	bottle bottle apple	*apple*	apple apple apple
1-05	color around red	*apple*	green pea yellow
2-00	eat apple red	*apple*	mm nice first
2-07	apple cut quarter	*apple*	seed pip core
3-00	thing type fruit	*apple*	pear orange share

Table 6.1: Examples of child-directed utterances for the target concepts *bed* and *apple* from the CHILDES corpus (after removal of stopwords and low frequency terms), together with the age of the addressed child.

two phenomena of meaning change: the development of featural concept representations in children (Section 6.3), and diachronic change of word meaning (Section 6.4).

6.3 Experiment 5: Development of Concept Representations in Infants

Children learn the meaning of concepts over time, and acquire increasingly nuanced and complex representations. We reviewed prior research in support of this claim in Section 6.1.1. To the best of our knowledge, we present the first computational investigation of this phenomenon at scale, modeling the development of representations comprising a broad variety of features for a large number of concepts. We model the development of concept representations by exposing SCAN, as introduced in Section 6.2, to a corpus of child-directed language.

We learn SCAN models for individual concepts (aka basic level categories such as *dog*, *chair*, or *ball*) from sets of input stimuli in the form of short child-directed text snippets comprising a mention of the target concept embedded in local context. Example documents are shown in Table 6.1. We model temporal meaning representations as a

	Full	**Thomas**
number of children	21	1
age range (y;mm)	0;11 – 4;11	2;00 – 4;11
number of utterances	129,958	45,081

Table 6.2: Details on the full corpus and the Thomas corpus. Note that the full corpus includes the Thomas corpus.

set of feature types. Our model captures (a) the internal development of feature types over time (for example a `color` feature type may contain increasingly nuanced color representations); and (b) the development of their relative importance (for example relational associations containing `travel`-related features may emerge over time in relation to *cars* or *trains*, gaining importance in relation to their perceivable features and leading to a more diverse concept representation).

In contrast to Chapters 4 and 5, which investigated the acquisition and representations of superordinate level categories, here we use our dynamic Bayesian model to study the meaning development of *basic level categories* (Rosch, 1978) in young infants. Our corpora comprise language directed to children from their first word onset (one year) up to about five years of age, and thus cover the initial phase of linguistic development. Basic level categories are used most frequently as labels by caretakers, and are the first categories children learn to distinguish (Rosch, 1978). Furthermore, basic level objects tend to be associated with a single word. Like in the previous models and experiments in this thesis, we treat a linguistic mention of a word referring to a target concept as an observation of the target concept itself, and its local context as the concept's features. We train one SCAN model per word (or concept) of interest.

The experiments presented in the following sections are designed to investigate quantitatively (Section 6.3.1) and qualitatively (Section 6.3.2) whether the representations that SCAN induces from corpora of child-directed language reflect characteristics of concept development in infants.

Data In order to capture change of meaning representations in children over time, we require longitudinal input data, i.e., (a) frequent recordings of language directed to the child (in order to learn time-specific meaning representations), and (b) recordings spanning a significant temporal period (in order to capture the development of these representations). In fact, the corpus used in the cognitive experiments in Chapter 5, derived from the CHILDES database of child-related speech (MacWhinney, 2000), was constructed with these desiderata in mind. The corpus is described in detail in Section 5.4 (page 145). From this underlying data set, we create two corpora of input stimuli for our experiments: one corpus conflates the data from the four sub-corpora (comprising input to 21 children). The other corpus contains only the Thomas corpus (Lieven et al., 2009), the largest longitudinal collection of input specific to one individual child. These two corpora allow us to investigate whether conflating data for many children, as opposed to data comprising input to only one child, has an influence on the model output. Details on the size and coverage of the corpora can be found in Table 6.2.

From each of the two data sets described above, we created concept-specific input for our SCAN model. We trained models for a set of 30 concepts, which are listed in Appendix C.1. The majority of this set (21 words) was taken from the data set of concepts of living and non-living things used in the evaluations in the preceding chapters (McRae et al., 2005; Vinson and Vigliocco, 2008), and described in detail in Section 4.4.1 (p. 80). We selected nouns based on a sufficient number of mentions in the child-directed data. In addition, we added verbs, superordinate categories and adjectives as target concepts, again selected based on their frequency in the data. The target concept-specific input corpora consist of short text snippets, containing a mention of the target concept surrounded by a symmetric window of $n = \pm 3$ content words (we remove stop words and low frequency terms). Each input is annotated with the age (in months) of the child being spoken to. Table 6.1 shows examples of input documents for the target concepts *bed* and *apple*.

6.3.1 Development of Feature Complexity

This experiment investigates the development of the complexity of concept-specific feature representations over time. We approximate the complexity of inferred featural representations through the age-of-acquisition (aoa) rating of their associated features (i.e., context words). Age-of-acquisition ratings measure the age at which a person understands the meaning of a word (but does not necessarily use it). Large databases of age-of-acquisition ratings exist that cover more than 30,000 English words (Kuperman et al., 2012). Age-of-acquisition has been shown to correlate with other measures of complexity, such as word length, concreteness, or imageability (Kuperman et al., 2012). We quantify the development of temporal interval-specific complexity of learnt representations as a function of the aoa scores of their associated features. We provide a *qualitative* analysis of the development of concept representations in Section 6.3.2.

We compare the difference in development of meaning representations between two models, trained on the respective corpora introduced above: The full corpus comprising input to 21 children, and the Thomas corpus containing input to a single child. This allows us to study whether the same pattern of feature development emerges for individuals, as well as across children.

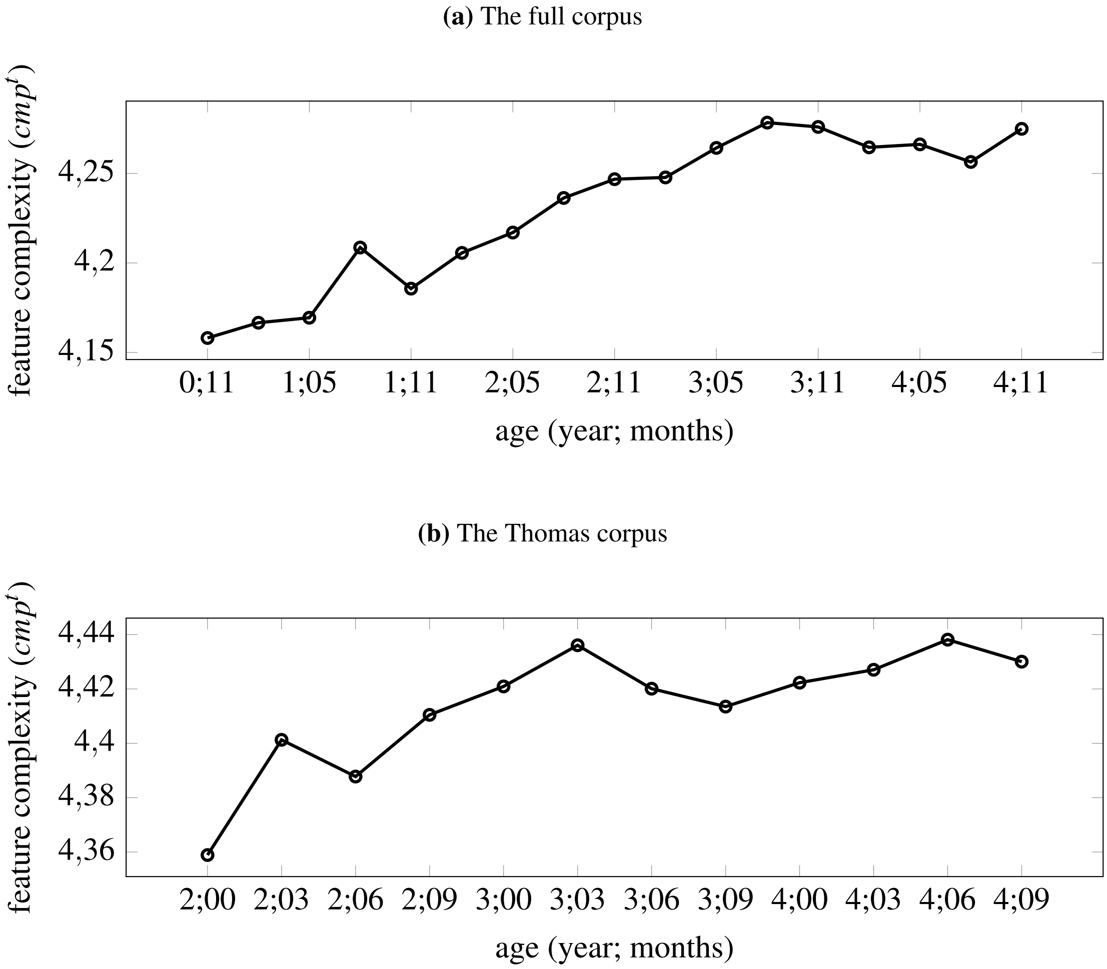
Models and parameters We create a SCAN model for each of the 30 concepts in our set of targets. SCAN models are parameterized with respect to the number of feature types (senses in the model description) they support, and with respect to the size of the temporal intervals. We set the number of feature types to $K = 5$, and the size of temporal intervals to $\Delta T = 3$ months. We set the word-feature type precision parameter $\kappa^{\Psi} = 50$ (triggering thematically stable feature types which refer to the same aspect of meaning across temporal intervals). We adjusted these parameters to the size and characteristics of our datasets using a small set of development concepts, but did not tune them exhaustively.

Method For each of the 30 target concepts, c, we induce time-specific concept representations as distributions over feature types g_t, where each feature type is represented as a distribution over features. We represent each induced feature type g_t as the 10 features with highest probability under g_t, $f_n^{g_t} : n = [1...10]$. Taken together for all 30 concepts, interval-specific feature sets comprise 1,500 context word token, and on average around 370 context word types (i.e., distinct features). For each feature in this set, we retrieve an age-of-acquisition rating $aoa(f_n^{g_t})$ from Kuperman et al. (2012)'s resource (which covers over 97% of the features in our data set).

We compute time interval t-specific complexity scores cmp^t by averaging feature aoa-scores over all target concepts c and all their time-specific feature types g_t,

$$cmp^t = \frac{1}{F} \sum_c \sum_{g_t} \sum_{n=1}^{10} aoa(f_n^{g_t}), \qquad (6.12)$$

where F is the total number of feature tokens in the time-specific representations.


(a) The full corpus

(b) The Thomas corpus

Figure 6.3: Development of complexity of featural concept representation with increasing age of children for the full corpus (top) and the Thomas corpus (bottom). Complexity is quantified through averaged age-of-acquisition of concept-associated features (cf., cmp^t, equation (6.12)).

Results Figure 6.3 displays the development of age-of-acquisition scores over time for both the full corpus (6.3a) and the child-specific Thomas corpus (6.3b). Across corpora, the average age-of-acquisition rating of features consistently increases. The trend is statistically significant (Spearman's $\rho = 0.91$ (full corpus), $\rho = 0.83$ (Thomas corpus); $p < 0.002$). Overall, the trend is more stable for the full corpus.

Note that SCAN is not a model of word learning – we use age-of-acquisition as a way to quantify the complexity of concept representations. The absolute values of the age-of-acquisition scores reported in Figures 6.3a and 6.3a do not correspond to the age of the child. Our model does not learn what these words mean, but it learns that they are features which are relevant to and representative of a concept. Based on repeatedly observed co-occurrences, a child may learn that certain words are associated with certain concepts without having a clear representation of their meaning.

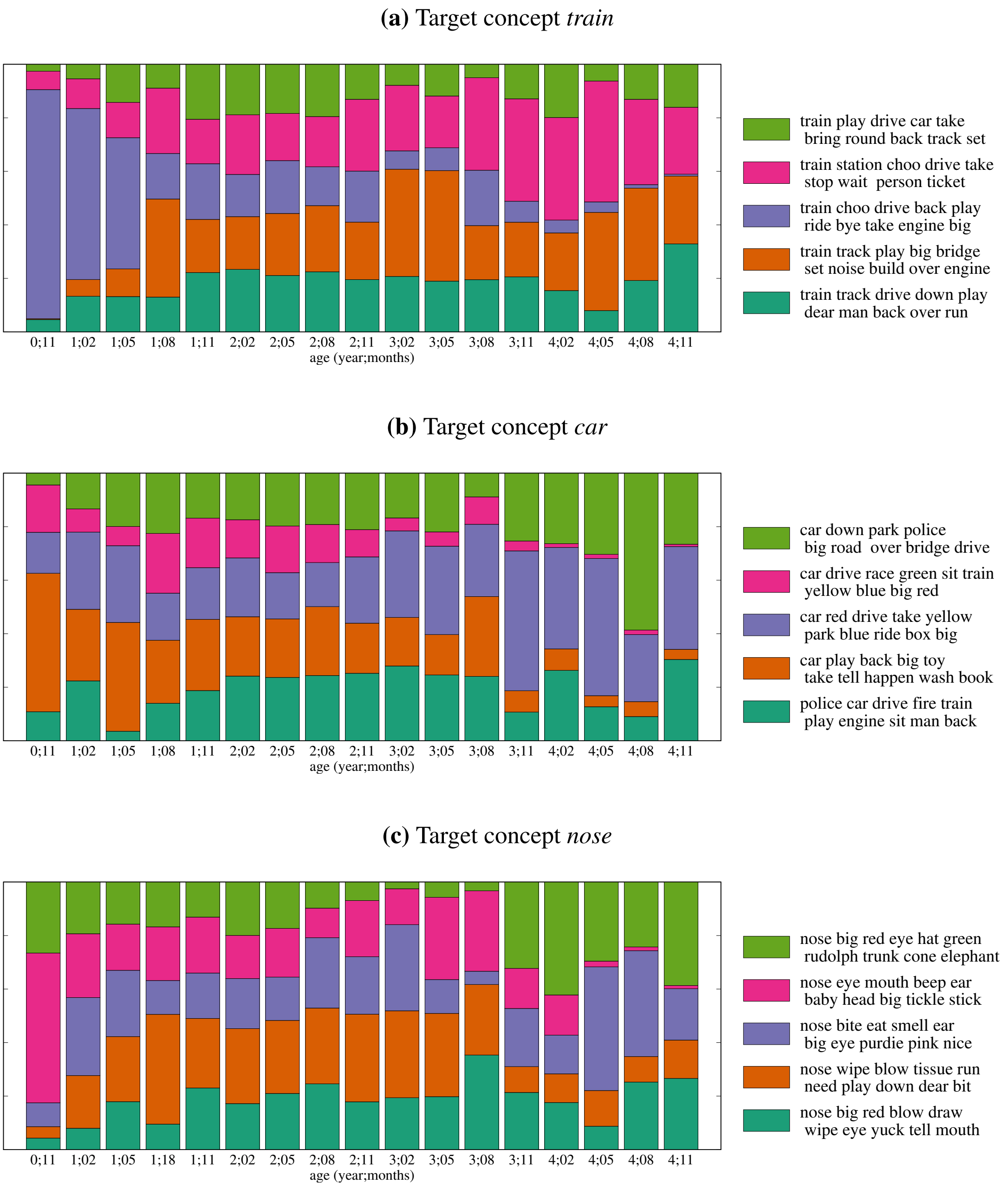
We provided quantitative support for the claim that our model learns concept representations that increase in complexity over time, mirroring the way in which children incrementally and dynamically acquire increasingly nuanced conceptual knowledge about the world. In the next experiment, we qualitatively analyze representations of a subset of these concepts as learnt by our model, and their qualitative dynamic development.

6.3.2 Qualitative Analysis of Feature Development

We present qualitative output of our SCAN models trained on a selection of target concepts, using the same set of models parameter settings as in the previous experiment. We compare differences in the development of meaning representations over time when training on the full corpus and the Thomas corpus. While we expect that the larger amount of training data available in the conflated corpus will trigger more stable representations, we also assume that time-specific representations are highly child-specific as they depend on the input and situations the child encounters. The representations induced from the Thomas corpus should reflect this.

Figures 6.4 and 6.5 display the development of meaning representations as captured by our model based on the full corpus of 21 children, and Figure 6.6 shows inferred representations from the Thomas corpus. Additional model output for both corpora is provided in Appendix C.2. Individual meaning representations are visualized as a bar capturing the relative prevalence $(p(k|t) = \phi^t_k)$ of different feature types (color-coded). One such visualization is displayed for each temporal interval, illustrating the development of feature type prevalence over time. Each feature type is illustrated to the right of the plot as the ten words w most highly associated with the feature type, marginalizing over the time-specific representations $\left(p(w|k) = \sum_t \psi^{t,k}_w\right)$.

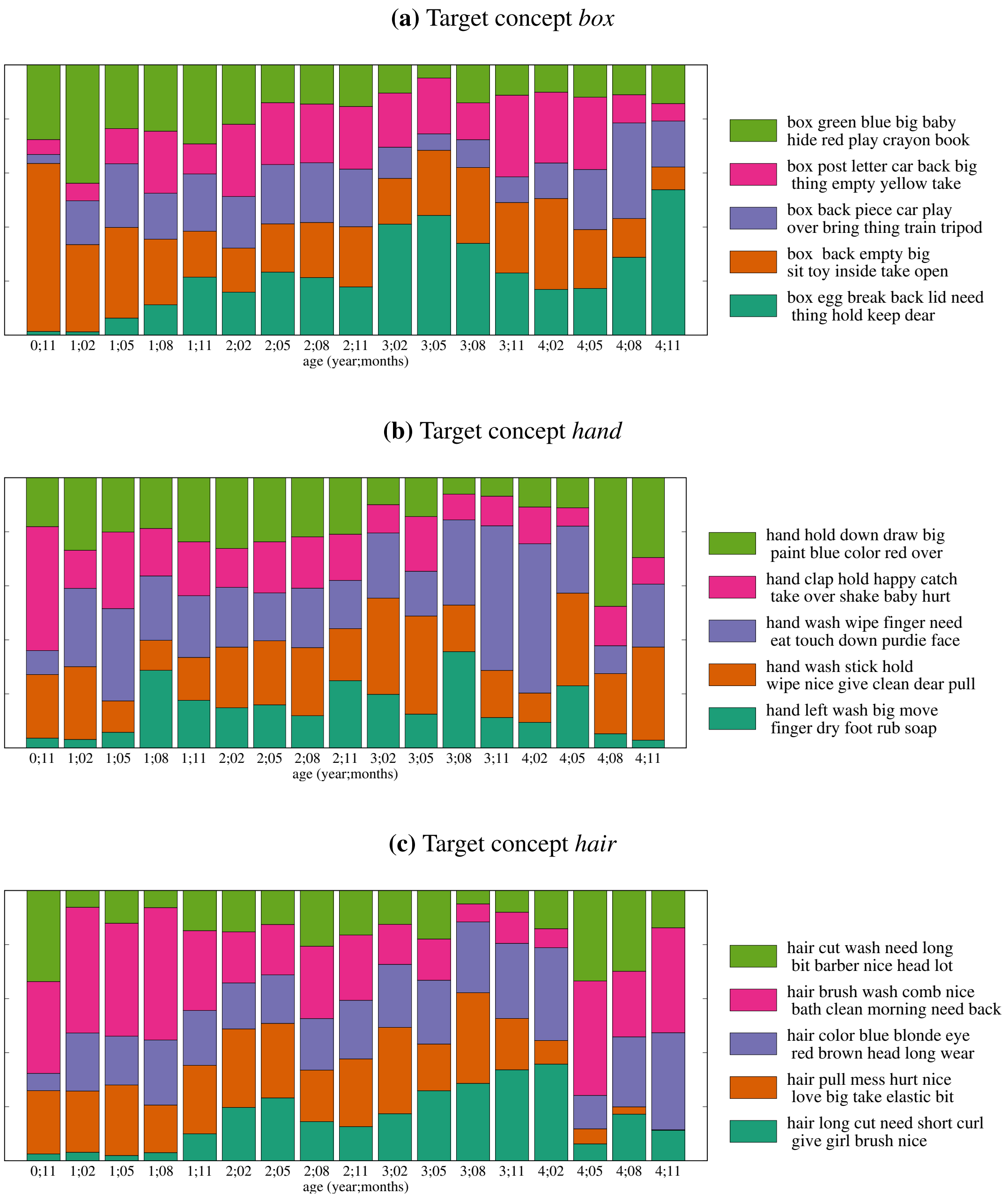
Analysis of the full corpus Figure 6.4a shows the meaning development of the concept *train*. Initially thematically rather unspecific feature type (violet) is prevalent. Over time features relating to `train journey` (pink; including words like {ticket, station, wait}), and `location` (orange; {bridge, track}) increase in importance, leading


(a) Target concept *train*

(b) Target concept *car*

(c) Target concept *nose*

Figure 6.4: Visualization of feature development of the concepts *train*, *car* and *nose* (top to bottom), based on the linguistic input to 21 children aged between 11 months and 4 years and 11 months. Each bar shows the proportional prevalence of each feature type (color-coded) and is labeled with the start year of the respective time interval (covering three months). Feature types are shown as the 10 most probable words to the right of the corresponding plot.


(**a**) Target concept *box*

(**b**) Target concept *hand*

(**c**) Target concept *hair*

Figure 6.5: Additional model output from the conflated corpus comprising input to 21 children for the feature development of the concepts *box*, *hand* and *hair* (top to bottom).

to a differentiation in meaning, and a more fine-grained representation of the concept *train*.

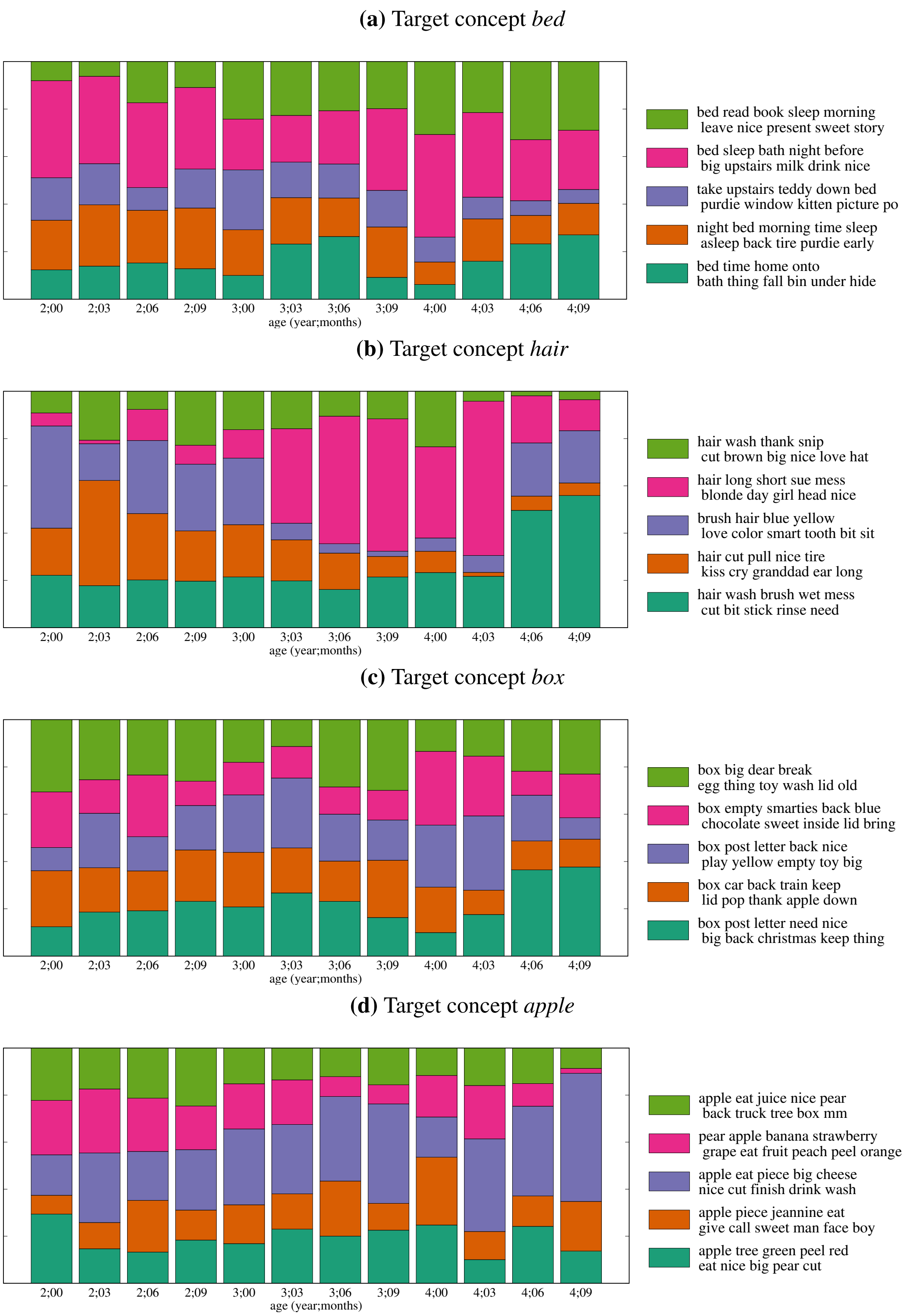
The graph in Figure 6.4b presents the temporal representations of the concept *car*. Mirroring the initial meaning of *train* discussed above, initially incoherent and vaguely `play`-related features dominate the representation (orange). Over time, features emphasizing conceptual associations increase in importance: the dark green and light green feature types cover concepts related to the target concept *car* (such as {road, police, bridge, engine}), leading to a more differentiated representation over time. The pink feature type captures `color` features.

We show the meaning development for the concept *nose* in Figure 6.4c. The initially prevalent pink feature type is very general comprising other bodyparts {eye, mouth} as well as related actions {tickle, stick}. The orange and dark green feature types which increase in prevalence over time focus on the `cleaning`-related associations of *nose*. The light green feature type is animal-related and suggests a broadening of associations to `animals` – featuring mentions of elephants and Rudolph (the reindeer).

The graph in Figure 6.5a presents the meaning development of the concept *box*. Once more, the initially prevalent feature type (orange) is topically incoherent. Over time, a `post box` association emerges (pink), as well as a feature type pertaining to `egg boxes` (dark green), together capturing a wider variety of specific aspects related to the concept *nose*.

Figure 6.5b presents the temporal representations of the concept *hand*. Initially the meaning is represented predominantly through one prevalent feature type (pink) which captures the general nature of speech directed to very small children ({clap, hold, happy, baby...}). Over time this feature type decreases in prevalence, making room for a `washing`-related feature type (orange/violet). The light green feature type relates to `painting` and indicates a development of associations with *hand*-related actions.

Analysis of the Thomas corpus Figure 6.6a presents the development of the representation of the concept *bed*. We can make out a feature type related to `sleeping`, `going-to-bed` (pink) which is prevalent throughout. A separate feature type covers the `waking up` aspect (orange). A `reading`-related feature type emerges from age of about 2.5 years (light green) and increases in prevalence throughout. Like the representations induced from the full corpus, the meaning representation becomes more



(a) Target concept *bed*

(b) Target concept *hair*

(c) Target concept *box*

(d) Target concept *apple*

Figure 6.6: Development of meaning representations of the concepts *bed*, *hair*, *box*, and *apple* (top to bottom). Representations are induced from the Thomas corpus capturing development between the age of 2 years and 4 years 11.

faceted over time.

The meaning development of the concept *hair* is displayed in the plot in Figure 6.6b. A {washing, cleaning}-related aspect of meaning increases in prevalence towards the end of the modeled period, suggesting a newly learnt association (dark green). The orange and pink feature types reveal the effect of modeling meaning representations based on input to only one child. Both refer to a particular person of Thomas' environment (his granddad and his friend Sue, respectively). For comparison, we show the meaning development for the concept *hair* learnt from the full corpus in Figure 6.5c. The learnt meaning aspects are more general.

The two bottom plots (Figures 6.6c and 6.6d) show the meaning development of concepts *box* and *apple*, respectively. The pink feature type of *box* shows another instantiation of individual differences in meaning representation, referring to smartie boxes. This type did not emerge from the model trained on the conflated data. For the word *apple* (Figure 6.6d) different aspects of its meaning clearly emerge: one corresponding to an apple as *food* (violet), one corresponding to its category fruit (pink), as well as an aspect corresponding to its natural origin (dark green).

6.3.3 Discussion

Children's representations of categories and concepts evolve over time until they reliably resemble the meanings which are shared in the society they grow up in. In this set of experiments we showed that dynamic development of concept meaning representations emerges from a computational model of concept acquisition from child-directed language. The linguistic contexts in which concepts occur in child-directed speech changes with increasing age of the child, and allows the acquisition of increasingly accurate and diverse meaning representations. We quantified the increasing complexity of concept representations by linking the induced feature types to the age of acquisition of their representative terms. In addition, we qualitatively analyzed the development of meaning representations. We observed the phenomenon of initial overgeneralization (cf., concepts *train* and *box* in Figures 6.4a and 6.5a), as well as a shift from child-like representations to more general representations (cf., concept *hand* in Figure 6.5b).

Recognizing that meaning development of concepts is highly individual and dependent on the child's personal environment (Neisser, 1987), we investigated concept development from input to a single child, and from input to multiple children. SCAN picks up meaning change in both settings. We found that the representations learnt from a single child's corpus are more individual (cf., concept *hair* in Figure 6.6b); and that the prevalence of these highly personalized aspects of meaning decrease over time, suggesting a generalization process. In future work it would be desirable to compare the individual differences in feature learning across multiple children. With the notable exception of the Thomas corpus used throughout our experiments, however, currently available data sets of speech directed to individual children are either sparse, providing only infrequent samples of recordings and/or cover a shorter time period which makes it difficult to detect feature development.

As in previous chapters of this thesis we model the acquisition and development of conceptual knowledge based on the linguistic environment. Is the development of concept representations induced by our model exclusively a by-product of conceptual development in the child's mind? Certainly there are factors beyond this development which lead to a qualitative change of the input data the child receives. Examples include developments in the child's behavior and abilities (e.g., the ability to use pens to paint pictures will influence the linguistic contexts in which a child observes the word *hand*) or changes in the child's general environment (e.g., interacting with unknown people or visiting novel places). Teasing apart changes in abilities and environment from children's conceptual development is challenging within our experimental setup, and provides an interesting direction for future investigations.

The number of feature types associated with a concept is a parameter of our model, and is constant over time. In this set of experiments we set this parameter to the same value for all concepts. A more realistic model should be able to (a) induce the number of feature types individually for each concept, and (b) within concepts allow this number to vary over time. While in principle our models can capture such trends by setting the number of feature types to a high value and letting the model decide to not make use of all feature types, a more principled model should be able to adapt in complexity as demanded by the data.

Our aim in this study was to show that the dynamic development of featural concept representations during language acquisition emerges in large-scale experiments based on naturalistic child-directed language. However, we do not claim to fully capture the meaning acquisition process with our model: Modeling featural development from linguistic input remains agnostic about pre-linguistic feature learning, e.g., from statistical regularities in visual input (Mervis, 1987; Younger and Fearing, 2000). For example, awareness of basic object properties, such as object permanence has been shown in 7 months old children (Baillargeon, 1987). Pre-linguistic development presumably influences subsequent language-based feature learning. Furthermore, by training concept-specific models, we assume that the learner has established an a priori one-to-one word-concept mapping. This assumption is crude because one word can map to a variety of concepts and vice versa. Furthermore, concept to word mappings are themselves *learnt* by children around the same time as conceptual representations are acquired (see our discussion in Section 2.1). We ignore information from the visual or pragmatic input available to the child, and our corpora only capture snapshots of specific situations the child encounters.

Despite these limitations, our text-based approach allows us to investigate the development of a broad class of features for a variety of concepts – their coverage being only restricted by the thematic variety in the corpus. Besides, previous analyses showed that language encodes a variety of information of other (e.g., visual) modalities, and that child-directed speech particularly often refers to perceivable properties of basic-level categories (Riordan and Jones, 2011; Callanan, 1990). We showed that patterns of child featural development identified in the literature emerge from our Bayesian models based on statistical regularities in the linguistic input to the child. In addition to incremental category learning (Chapter 4) and joint acquisition of categories and features (Chapter 5) the phenomenon of dynamic meaning acquisition can be successfully captured with a Bayesian model trained on corpora of child-directed language.

6.4 Experiment 6: Development of Word Meaning

This section evaluates SCAN's ability to capture phenomena related to diachronic meaning change. Evaluation of models which detect meaning change is fraught with difficulties. There is no standard set of words which have undergone meaning change or benchmark corpus which represents a variety of time intervals and genres, and is thematically consistent. Previous work has generally focused on a few hand-selected words and models were evaluated qualitatively by inspecting their output, or the extent to which they can detect meaning changes from two time periods. For example, Cook et al. (2014) manually identify 13 target words which undergo meaning change in a focus corpus with respect to a reference corpus (both news text). They then assess how their models fare at learning sense differences for these targets compared to distractors which did not undergo meaning change. They also underline the importance of using thematically comparable reference and focus corpora to avoid spurious differences in word representations.

We evaluate our model's ability to detect and quantify meaning change across several time intervals (not just two). Instead of relying on a few hand-selected target words, we use larger sets sampled from our learning corpus or found to undergo meaning change in a judgment elicitation study (Gulordava and Baroni 2011, Section 6.4.2). In addition, we adopt the evaluation paradigm of (Mitra et al. 2014, Section 6.4.3) and validate our findings against WordNet. Finally, we apply our model to the recently established SemEval-2015 diachronic text evaluation subtasks (Popescu and Strapparava 2015, Section 6.4.4). In order to present a consistent set of experiments, we use our own corpus throughout which covers a wider range of time intervals and is compiled from a variety of genres and sources and is thus thematically coherent (and described in detail below). Wherever possible, we compare against prior art, with the caveat that the use of a different underlying corpus unavoidably influences the obtained semantic representations.

Data The corpus described in the following underlies all experiments described in this section. We created a DiAchronic TExt corpus (DATE) which collates documents spanning years 1700–2010 from three sources: (a) the COHA corpus[10] (Davies, 2010), a large collection of texts from various genres covering the years 1810–2010; (b) the training data provided by the DTE task[11] organizers (see Section 6.4.4); and (c) the portion of the CLMET3.0[12] corpus (Diller et al., 2011) corresponding to the period 1710–1810 (which is not covered by the COHA corpus and thus underrepresented in our training data). CLMET3.0 contains texts representative of a range of genres including narrative fiction, drama, letters, and was collected from various online archives. Table 6.3 provides details on the size of our corpus. Documents were clustered by their year of publication as indicated in the original corpora. In the CLMET3.0 corpus, occasionally a range of years would be provided. In this case we used the final year of the range. We tokenized, lemmatized, and part of speech tagged DATE using the NLTK (Bird et al., 2009). We removed stopwords and function words. After preprocessing,

[10] http://corpus.byu.edu/coha/

[11] http://alt.qcri.org/semeval2015/task7/index.php?id=data-and-tools

[12] http://www.kuleuven.be/~u0044428/clmet3_0.htm

Corpus	years covered	#words
COHA	1810–2009	142,587,656
DTE	1700–2010	124,771
CLMET3.0	1710–1810	4,531,505

Table 6.3: Size and coverage of our three training corpora (after pre-processing).

year	text snippet		
1700	ambassador emperor treat peace king	*power*	enlarge english slave dominion condition
1838	sharp listen noble school awaken	*power*	mind exercise wit head house
1867	drainage wit rapidity flow water	*power*	remove obstacle practice stream wend
1989	governmental action individual equal	*power*	preponderant force energy direction govern
2010	invest million dollar building thermal	*power*	plant bid tide crisis brazilian

Table 6.4: Example text snippets for the target concept *power* from our DATE corpus (after removal of stopwords and low frequency terms), together with the respective year of origin.

we extracted target word-specific input corpora for our models. These consisted of mentions of a target c and its surrounding context, a symmetric window of ± 5 words. Example documents for the target word *power* are displayed in Table 6.4.

6.4.1 Temporal Dynamics

As discussed in Section 6.1.2.1, our model departs from previous approaches (e.g., Mitra et al. 2014) in that it learns globally consistent temporal representations for each word. In order to assess whether temporal dependencies are indeed beneficial, we implemented a stripped-down version of our model (SCAN-NOT) which does not have any temporal dependencies between individual time steps (i.e., without the chain iGMRF priors). Word meaning is still represented as senses and sense prevalence is modeled as a distribution over senses for each time interval. However, time intervals are now independent. Inference works as described in Section 6.2.2, without having to learn the κ precision parameters.

Models and Parameters We compared the two models in terms of their predictive power. We split the DATE corpus into a training period $\{d^1...d^t\}$ of time slices 1 through t and computed the likelihood $p(d^{t+1}|\phi^t, \psi^t)$ of the data at test time slice $t+1$, under the parameters inferred for the previous time slice. The time slice size was set to $\Delta T = 20$ years. We set the number of senses to $K = 8$, the word precision parameter $\kappa^\psi = 10$, a high value which triggers individual senses to remain thematically consistent over time. We set the initial sense precision parameter $\kappa^\phi = 4$, and the Gamma parameters $a = 7$ and $b = 3$. These parameters were optimized once on the development data used for the task-based evaluation discussed in Section 6.4.4. Unless otherwise specified all experiments reported in this section use these values. No parameters were tuned on the test set for any task. In all experiments we ran the Gibbs sampler for 1,000 iterations, and resampled κ^ϕ after every 50 iterations, starting from iteration 150. We report results based on the final state of the sampler throughout. We randomly selected 50 mid-frequency target concepts from a larger set of target concepts described in Section 6.4.4. Predictive log-likelihood scores were averaged across concepts and were calculated as the average under 10 parameter samples $\{\phi^t, \psi^t\}$ from the trained models.

Results Figure 6.7 displays predictive log-likelihood scores for four test time intervals. SCAN outperforms its stripped-down version throughout (higher is better). Since the representations learnt by SCAN are influenced (or smoothed) by neighboring representations, they overfit specific time intervals less which leads to better predictive performance.

Figure 6.8 further illustrates how SCAN captures meaning change for the words *band*, *power*, *transport* and *bank*. The sense distributions over time are shown as a sequence of stacked histograms, senses themselves are color-coded (and enumerated) below, in the same order as in the histograms. Each sense k is illustrated as the 10 words w assigned the highest posterior probability, marginalizing over the time-specific representations $p(w|k) = \sum_t \psi_w^{t,k}$. Words representative of prevalent senses are highlighted in bold face.

Figure 6.8a demonstrates that the model is able to capture various senses of the word *band*, such as `strip used for binding` (yellow bars/number 3 in the figure) or `musical band` (grey/1, orange/7). Our model predicts an increase in prevalence over the modeled time period for both senses. This is corroborated by the OED which provides the majority of references for the `binding strip` sense for the 20th century and dates the `musical band` sense to 1812. In addition a `social band` sense (violet/6,

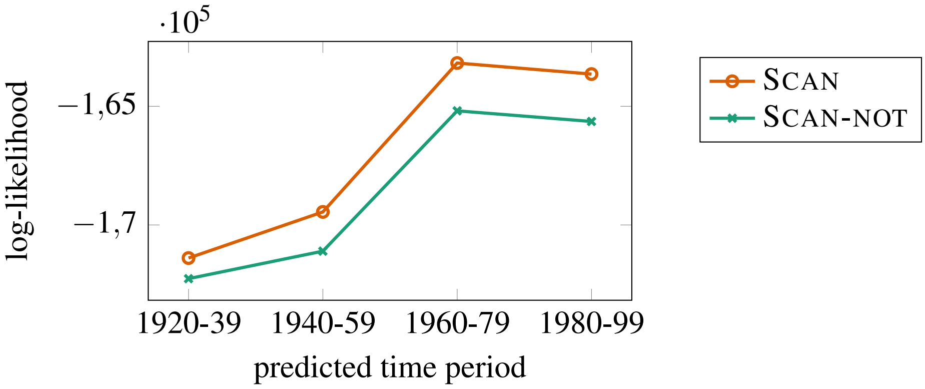


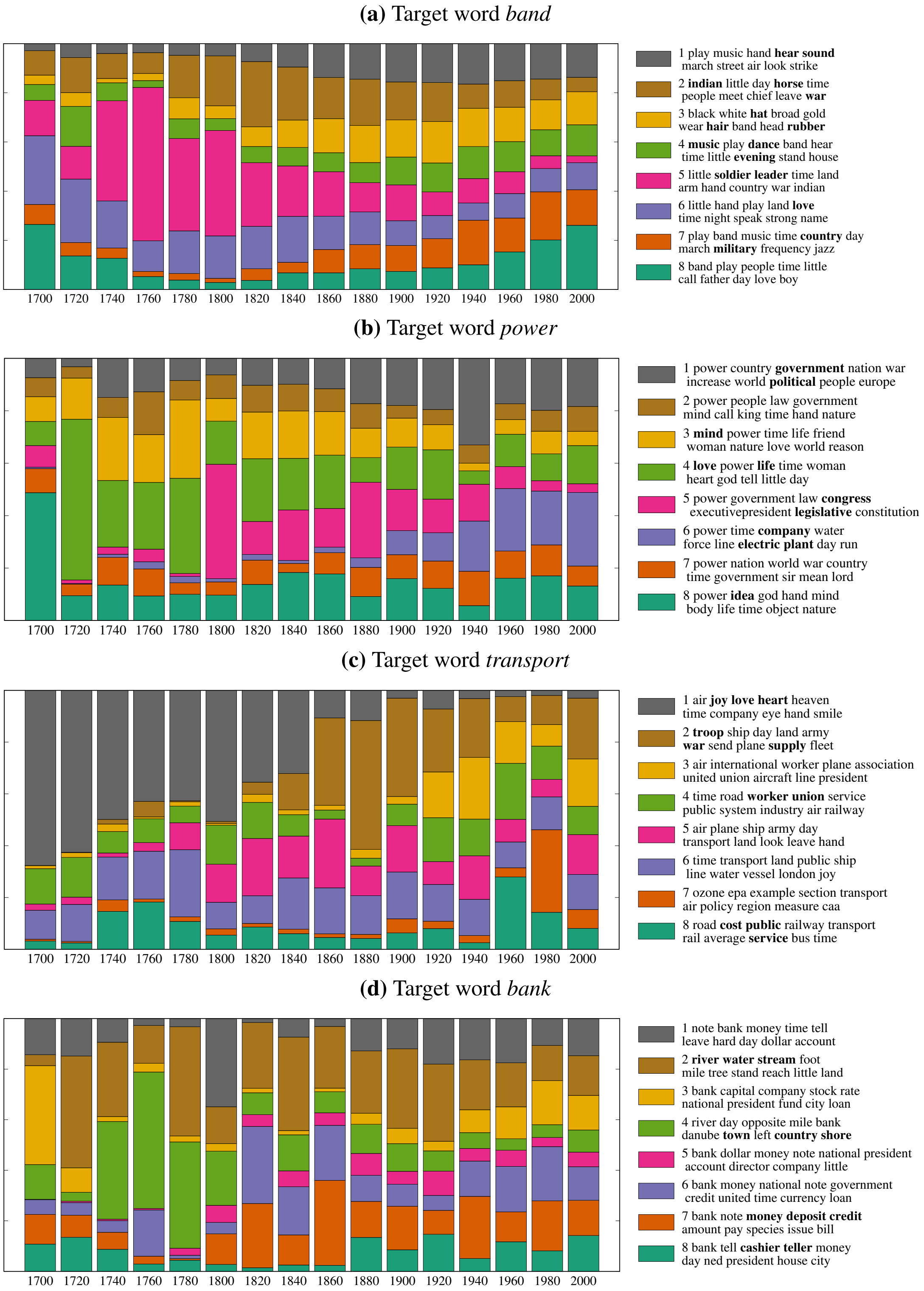
Figure 6.7: Predictive log likelihood of SCAN and a version without temporal dependencies (SCAN-NOT) across various test time periods.

darkgreen/8; in the sense of bonding) emerges, which is present across time slices. The sense colored brown/2 refers to the `British Band`, a group of native Americans involved in the Black Hawk War in 1832, and the model indeed indicates a prevalence of this sense around this time (see bars 1800–1840 in the figure).

For the word *power* (Figure 6.8b), three senses emerge: the `institutional power` (colors gray/1, brown/2, pink/5, orange/7 in the figure), `mental power` (yellow/3, lightgreen/4, darkgreen/8), and `power as supply of energy` (violet/6). The latter is an example of a "sense birth" (Mitra et al., 2014): the sense was hardly present before the mid-19th century. This is corroborated by the OED which dates the sense to 1889, whereas the OED contains references to the remaining senses for the whole modeled time period, as predicted by our model.

Similar trends of meaning change emerge for *transport* (Figure 6.8c). The plot in Figure 6.8d shows the sense development for the word *bank*. Although the well-known senses `river bank` (brown/2, lightgreen/4) and `monetary institution` (rest) emerge clearly, the overall sense pattern appears comparatively stable across intervals indicating that the meaning of the word has not changed much over time.

Besides tracking sense prevalence over time, our model can also detect changes within individual senses. Because we are interested in tracking semantically stable senses, we fixed the precision parameter κ^{Ψ} to a high value, to discourage too much variance within each sense. Figure 6.9 illustrates how the `energy` sense of the word *power* (violet/6 in Figure 6.8) has changed over time. Selected characteristic terms are highlighted in bold face. For example, the term "water" is initially prevalent, while the term "steam" rises in prevalence towards the middle of the modeled period, and is



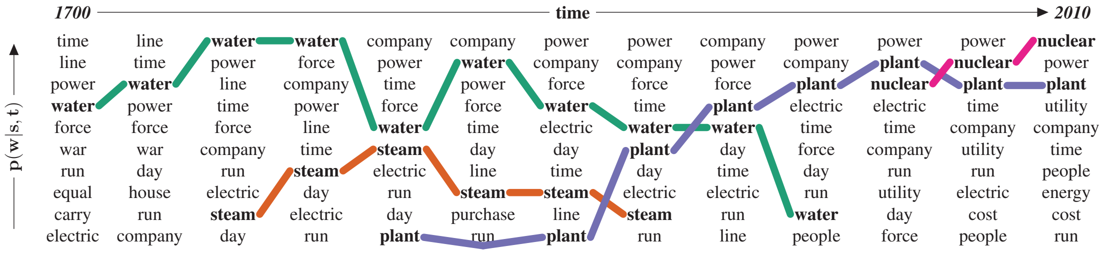
(a) Target word *band*

(b) Target word *power*

(c) Target word *transport*

(d) Target word *bank*

Figure 6.8: Tracking meaning change for the words *band*, *power*, *transport* and *bank* over 20-year time intervals between 1700 and 2010. Each bar shows the proportion of each sense (color-coded) and is labeled with the start year of the respective time interval. Senses are shown as the 10 most probable words, and particularly representative words are highlighted for illustration.



time	line	**water**	**water**	company	company	power	power	company	power	power	power	**nuclear**
line	time	power	force	power	**water**	company	company	power	company	**plant**	**nuclear**	power
power	**water**	line	company	time	power	force	force	force	**plant**	**nuclear**	**plant**	**plant**
water	power	time	power	force	force	**water**	time	**plant**	electric	electric	time	utility
force	force	force	line	**water**	time	electric	**water**	**water**	time	time	company	company
war	war	company	time	**steam**	day	day	**plant**	day	force	company	utility	time
run	day	run	**steam**	electric	line	time	day	time	day	run	run	people
equal	house	electric	day	run	**steam**	**steam**	electric	electric	run	utility	electric	energy
carry	run	**steam**	electric	day	purchase	line	**steam**	run	**water**	day	cost	cost
electric	company	day	run	**plant**	run	**plant**	run	line	people	force	people	run

Figure 6.9: Sense-internal temporal dynamics for the `energy` sense of the word *power* (violet/6 in Figure 6.8). Columns show the ten most highly associated words for each time interval for the period between 1700 and 2010 (ordered by decreasing probability). We highlight how four terms characteristic of the sense develop over time ({water, steam, plant, nuclear}).

superseded by the terms “plant” and “nuclear” towards the end.

6.4.2 Novel Word Sense Detection

In this section and the next we will explicitly evaluate the temporal representations (i.e., probability distributions) induced by our model, and discuss its performance in the context of previous work.

Large-scale evaluation of meaning change is notoriously difficult, and many evaluations are based on small hand-annotated goldstandard data sets. Mitra et al. (2015), bypass this issue by evaluating the output of their system against WordNet (Fellbaum, 1998a). Here, we consider their automatic evaluation of sense-births, i.e., the emergence of novel senses. We assume that novel senses are detected at a focus time t_2 whilst being compared to a reference time t_1. WordNet is used to confirm that the proposed novel sense is indeed distinct from all other induced senses for a given word.

Method Mitra et al.’s (2015) evaluation method presupposes a system which is able to detect senses for a set of target words and identify which ones are novel. Our model does not automatically yield novelty scores for the induced senses. However, Cook et al. (2014) propose several ways to perform this task post-hoc. We use their *relevance* score, which is based on the intuition that keywords (or collocations) which characterize the difference of a focus corpus from a reference corpus are indicative of word sense novelty.

We identify keywords for a focus corpus with respect to a reference corpus using Kilgarriff's (2009) method which is based on smoothed relative frequencies.[13] The novelty of an induced sense s can be then defined in terms of the aggregate keyword probabilities given that sense (and focus time of interest):

$$rel(s) = \sum_{w \in W} p(w|s,t_2). \tag{6.13}$$

where W is a keyword list and t_2 the focus time. Cook et al. (2014) suggest a straightforward extrapolation from sense novelty to word novelty:

$$rel(c) = \max_{s} \; rel(s), \tag{6.14}$$

where $rel(c)$ is the highest novelty score assigned to any of the target word's senses. A high $rel(c)$ score suggests that a word has undergone meaning change.

We obtained candidate terms and their associated novel senses from the DATE corpus, using the *relevance* metric described above. The novel senses from the focus period and all senses induced for the reference period, except for the one corresponding to the novel sense, were passed on to Mitra et al.'s (2015) WordNet-based evaluator which proceeds as follows. Firstly, each induced sense s is mapped to the WordNet synset u with the maximum overlap:

$$synset(s) = \arg\max_{u} \; overlap(s,u). \tag{6.15}$$

Next, a predicted novel sense n is deemed truly novel if its mapped synset is distinct from any synset mapped to a different induced sense:

$$\forall_{s'} synset(s') \neq synset(n). \tag{6.16}$$

Finally, overall precision is calculated as the fraction of sense-births confirmed by WordNet over all birth-candidates proposed by the model. Like Mitra et al. (2015) we only report results on target words for which all induced senses could be successfully mapped to a synset.

Models and Parameters We obtained the broad set of target words used for the task-based evaluation (in Section 6.4.4) and trained models on the DATE corpus. We set the number of senses $K = 4$ following Mitra et al. (2015) who note that the WordNet mapper works best for words with a small number of senses, and the time intervals to $\Delta T = 20$ as in the previous experiment. We identified 200 words[14] with highest nov-

[13] We set the smoothing parameter to $n = 10$, and like Cook et al. (2014) retrieve the top 1000 keywords.

[14] This threshold was tuned on one reference-focus time pair.

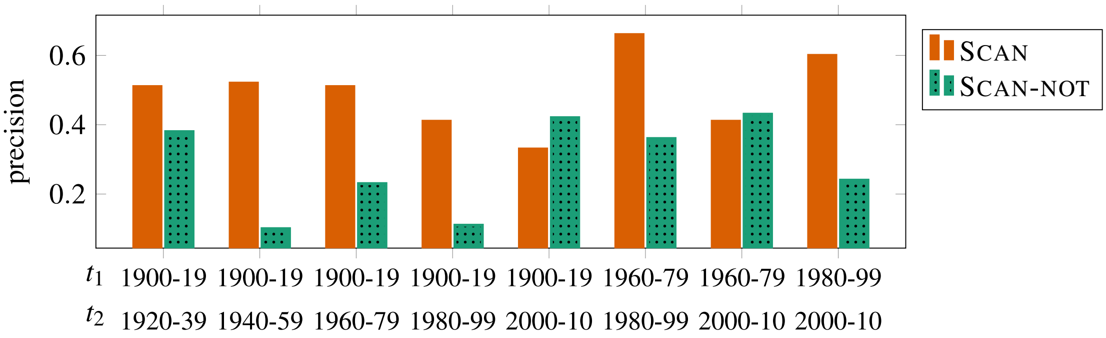


Figure 6.10: Precision results for the SCAN and SCAN-NOT models on the WordNet-based novel sense detection. Results are shown for a selection of reference times (t_1) and focus times (t_2).

elty score (Equation (6.14)) as sense birth candidates. We compared the performance of the full SCAN model against SCAN-NOT which learns senses independently for time intervals. We trained both models on the same data with identical parameters. For SCAN-NOT, we must post-hoc identify corresponding senses across time intervals. We used the Jensen-Shannon divergence between the reference- and focus time-specific word distributions $JS(p(w|s,t_1)||p(w|s,t_2))$ and assigned each focus-time sense to the sense with smallest divergence at reference time.

Results Figure 6.10 shows the performance of our models on the task of sense birth detection. SCAN performs better than SCAN-NOT, underscoring the importance of joint modeling of senses across time slices and incorporation of temporal dynamics. Our accuracy scores are in the same ballpark as Mitra et al. (2014, 2015). Note, however that the scores are not directly comparable due to differences in training corpora, focus and reference times, and candidate words. Mitra et al. (2015) use the larger Google syntactic n-gram corpus, as well as richer linguistic information in terms of syntactic dependencies. We show that our model which does not rely on syntactic annotations performs competitively even when trained on smaller data. Table 6.5 (top) displays examples of words assigned highest novelty scores for the reference period 1900–1919 and focus period 1980–1999, as induced by SCAN models.

	t_1=1900–1919 t_2=1980–1999
union	soviet united american union european war civil military people liberty
dos	system window disk pc operate program run computer de dos
entertainment	television industry program time business people world president company
station	radio station television local program network space tv broadcast air
	t_1=1960–1969 t_2=1990–1999
environmental	supra note law protection id agency impact policy factor federal
users	computer window information software system wireless drive web building
virtual	reality virtual computer center experience week community separation
disk	hard disk drive program computer file store ram business embolden

Table 6.5: Example target terms (left) with novel senses (right) as identified by SCAN in focus corpus t_2 (when compared against reference corpus t_1). Top: terms used in novel sense detection study (Section 6.4.2). Bottom: terms from the Gulordava and Baroni (2011) gold standard of word meaning change (Section 6.4.3).

6.4.3 Word Meaning Change

In this experiment we evaluate whether model induced temporal word representations capture perceived word novelty. We adopt the evaluation framework (and data set) introduced in Gulordava and Baroni (2011).[15]

Method Gulordava and Baroni (2011) do not model word senses directly; instead they obtain distributional representations of words from the Google Books (bigram) data for two time slices, namely the 1960s (reference corpus) and 1990s (focus corpus). To detect change in meaning, they measure cosine similarity between the vector representations of a target word in the reference and focus corpus. It is assumed that low similarity indicates that a word has undergone meaning change. To evaluate the output of their system, they created a test set of 100 target words (nouns, verbs, and adjectives), and asked five annotators to rate each word with respect to its degree of meaning change between the 1960s and the 1990s. The annotators used a 4-point ordinal scale (0: no change, 1: almost no change, 2: somewhat change, 3: changed significantly). Words were subsequently ranked according to the mean rating given by the annotators. Inter-annotator agreement on the novel sense detection task was 0.51

[15] We thank Kristina Gulordava for sharing their evaluation data set of target words and human judgments.

system	corpus	Spearman's ρ
Gulordava (2011)	Google	0.386
SCAN	DATE	0.377
SCAN-NOT	DATE	0.255
frequency baseline	DATE	0.325

Table 6.6: Spearman's ρ rank correlations between system novelty rankings and the human-produced ratings. All correlations are statistically significant ($p < 0.02$). Results for SCAN and SCAN-NOT are averages over five trained models.

(pairwise Pearson correlation) and can be regarded as an upper bound on model performance.

Models and Parameters We trained SCAN models for all words in Gulordava and Baroni's (2011) goldstandard. We used the DATE subcorpus covering years 1960 through 1999 partitioned by decade ($\Delta T = 10$). The first and last time interval were defined as reference and focus time, respectively (t_1=1960–1969, t_2=1990–1999). As in the previous experiment, a novelty score was assigned to each target word (using Equation (6.14)). We computed Spearman's ρ rank correlations between gold standard and model rankings (Gulordava and Baroni, 2011). We trained SCAN models setting the number of senses to $K = 8$. We also trained SCAN-NOT models with identical parameters. We report results averaged over five independent parameter estimates. Finally, as in Gulordava and Baroni (2011) we compare against a frequency baseline which ranks words by their log relative frequency in the reference and focus corpus.

Results The results of this evaluation are shown in Table 6.6. As can be seen, SCAN outperforms SCAN-NOT and the frequency baseline. For reference, we also report the correlation coefficient obtained in Gulordava and Baroni (2011) but emphasize that the scores are not directly comparable due to differences in training data: Gulordava and Baroni (2011) use the Google bigrams corpus (which is much larger compared to DATE). Table 6.5 (bottom) displays examples of words which achieved highest novelty scores in this evaluation and their associated novel senses, as induced by SCAN models.

6.4.4 Diachronic Text Classification

In the previous sections we demonstrated how SCAN captures meaning change between two periods. In this section, we assess our model on an extrinsic task which relies on meaning representations spanning several time slices. We quantitatively evaluate our model on the SemEval-2015 benchmark data sets released as part of the Diachronic Text Evaluation exercise (Popescu and Strapparava 2015; DTE). In the following we first present the DTE subtasks, and then describe our experimental setup, and systems used for comparison to our model.

SemEval DTE Tasks Diachronic text evaluation is an umbrella term used by the SemEval-2015 organizers to represent three subtasks aiming to assess the performance of computational methods used to identify when a piece of text was written. A similar problem is tackled in Chambers (2012) who label documents with time stamps whilst focusing on explicit time expressions and their discriminatory power. The SemEval data consists of news snippets, which range between a few words and multiple sentences. A set of training snippets, as well as gold-annotated development and test data sets are provided. DTE subtasks 1 and 2 involve temporal classification: given a news snippet and a set of non-overlapping time intervals covering the period 1700 through 2010, the system's task is to select the interval corresponding to the snippet's year of origin. Temporal intervals are consecutive and constructed such that the correct interval is centered around the actual year of origin. For both tasks temporal intervals are created at three levels of granularity (fine, medium, and coarse).

Subtask 1 involves snippets which contain an explicit cue for time of origin. The presence of a temporal cue was determined by the organizers by checking the entities' informativeness in external resources. Consider the example below:

(6.17) President de Gaulle favors an independent European nuclear striking force

The mentions of French president de Gaulle and nuclear warfare suggest that the snippet was written after the mid-1950s and indeed it was published in 1962. A hypothetical system would then have to decide amongst the following classes:

$$\{1700\text{–}1702, 1703\text{–}1705, \ldots, 1961\text{–}1963, \ldots, 2012\text{–}2014\}$$
$$\{1699\text{–}1706, 1707\text{–}1713, \ldots, 1959\text{–}1965, \ldots, 2008\text{–}2014\}$$
$$\{1696\text{–}1708, 1709\text{–}1721, \ldots, 1956\text{–}1968, \ldots, 2008\text{–}2020\}$$

The first set of classes correspond to fine-grained intervals of 2-years, the second set to medium-grained intervals of 6-years and the third set to coarse-grained intervals of 12-years. For the snippet in example (6.17) classes 1961–1963, 1959–1965, and 1956–1968 are the correct ones.

Subtask 2 involves temporal classification of snippets which lack explicit temporal cues, but contain implicit ones, e.g., as indicated by lexical choice or spelling. The snippet in example (6.18) was published in 1891 and the spelling of *to-day*, which was common up to the early 20th century, is an implicit cue:

(6.18) The local wheat market was not quite so strong to-day as yesterday.

Like in subtask 1, systems select a temporal interval from a set of contiguous time intervals of differing granularity. For this task, which is admittedly harder, levels of temporal granularity are coarser corresponding to 6-, 12- and 20-year intervals.

Participating SemEval Systems We compared our model against three other systems which participated in the SemEval task.[16] AMBRA (Zampieri et al., 2015) adopts a learning-to-rank modeling approach and uses several stylistic, grammatical, and lexical features. IXA (Salaberri et al., 2015) uses a combination of approaches to determine the period of time in which a piece of news was written. This involves searching for specific mentions of time within the text, searching for named entities present in the text and then establishing their reference time by linking these to Wikipedia, using Google n-grams, and linguistic features indicative of language change. Finally, UCD (Szymanski and Lynch, 2015) employs SVMs for classification using a variety of informative features (e.g., POS-tag n-grams, syntactic phrases), which were optimized for the task through automatic feature selection.

Models and Parameters We trained our model for individual words and obtained representations of their meaning for different points in time. Our set of target words consisted of all nouns which occurred in the development data sets for DTE subtasks 1 and 2 as well as all verbs which occurred at least twice in this data set. After removing infrequent words we were left with 883 words (out of 1,116). Target words were not optimized with respect to the test data in any way; it is thus reasonable to expect better performance with an adjusted set of words.

[16] We do not report results for the system USAAR which achieved close to 100% accuracy by searching for the test snippets on the web, without performing any temporal inference.

We set the model time interval to $\Delta T = 5$ years and the number of senses per word to $K = 8$. We also evaluated SCAN-NOT, the stripped-down version of SCAN, with identical parameters. Both SCAN and SCAN-NOT predict the time of origin for a test snippet as follows. We first detect mentions of target words in the snippet. Then, for each mention c we construct a document, akin to the training documents, consisting of c and its context $\mathbf{w}$, the ±5 words surrounding c. Given $\{c, \mathbf{w}\}$, we approximate a distribution over time intervals as:

$$p^{(c)}(t|\mathbf{w}) \propto p^{(c)}(\mathbf{w}|t) \times p^{(c)}(t). \tag{6.19}$$

The superscript (c) denotes parameters from the word-specific model. We marginalize over senses and assume a uniform distribution over time slices $p^{(c)}(t)$. Finally, we combine the word-wise predictions into a final distribution $p(t) = \prod_c p^{(c)}(t|, \mathbf{w})$, and predict the time t with highest probability.

Supervised Classification We also apply our model in a supervised setting, i.e., by extracting features for classifier prediction. Specifically, we trained a multiclass SVM (Chang and Lin, 2011) on the training data provided by the SemEval organizers (for DTE tasks 1 and 2). For each observed word within each snippet, we added as feature its most likely sense k given t, the true time of origin:

$$\arg\max_k \ p^{(c)}(k|t). \tag{6.20}$$

We also trained a multiclass SVM using character n-gram ($n \in \{1,2,3\}$) features in addition to the model features. Szymanski and Lynch (2015) identified character n-grams as the most predictive feature for temporal text classification using SVMs. Their system (UCD) achieved the best published scores in DTE subtask 2. Following their approach, we included all n-grams that were observed more than 20 times in the DTE training data.

Results We employed two evaluation measures proposed by the DTE organizers. These are precision p, i.e., the percentage of times a system has predicted the correct time period. And accuracy acc which is more lenient, and penalizes system predictions proportional to their distance from the true interval. We compute the p and acc scores for our models using the evaluation script provided by the SemEval organizers. Table 6.7 summarizes our results for DTE subtasks 1 and 2. We compare SCAN

	Task 1					
	2 yr		6 yr		12 yr	
	acc	*p*	*acc*	*p*	*acc*	*p*
Baseline	.097	.010	.214	.017	.383	.046
SCAN-NOT	.265	**.086**	.435	**.139**	.609	.169
SCAN	**.353**	.049	**.569**	.112	**.748**	**.206**
IXA	.187	.020	.375	.041	.557	.090
AMBRA	.167	.037	.367	.071	.554	.074
SVM SCAN	.192	.034	.417	.097	.545	.127
SVM SCAN+ngram	.222	.030	.467	.079	.627	.142

	Task 2					
	6 yr		12 yr		20 yr	
	acc	*p*	*acc*	*p*	*acc*	*p*
Baseline	.199	.025	.343	.047	.499	.057
SCAN-NOT	.259	.041	.403	.056	.567	.098
SCAN	.376	.053	.572	.091	.719	.135
IXA	.261	.037	.428	.067	.622	.098
AMBRA	.605	.143	.767	.143	.868	.292
UCD	**.759**	.463	**.846**	.472	**.910**	.542
SVM SCAN	.573	.331	.667	.368	.790	.428
SVM SCAN+ngram	.747	**.481**	.821	**.500**	.897	**.569**

Table 6.7: Results on Diachronic Text Evaluation Tasks 1 and 2 for a random baseline, our SCAN model, its stripped-down version without iGMRFs (SCAN-NOT), the SemEval submissions (IXA, AMBRA and UCD), and SVMs trained with SCAN features (SVM SCAN), and with additional character n-gram features (SVM SCAN+ngram). Results are shown for three levels of granularity, a strict precision measure p, and a distance-discounting measure acc.

against a baseline which selects a time interval at random[17] averaged over five runs. We also show results for a stripped-down version of our model without the iGMRFs (SCAN-NOT) and for the systems which participated in SemEval.

[17] We recomputed the baseline scores for subtasks 1 and 2 due to inconsistencies in the results provided by the DTE organizers.

For subtask 1, the two versions of SCAN outperform all SemEval systems across the board. SCAN-NOT occasionally outperforms SCAN in the strict precision metric, however, the full SCAN model consistently achieves better accuracy scores which are more representative since they factor in the proximity of the prediction to the true value. In subtask 2, the UCD and SVM SCAN+ngram systems perform comparably. They both use SVMs for the classification task, however our own model employs a less expressive feature set based on SCAN and character n-grams, and does not take advantage of feature selection which would presumably enhance performance. With the exception of AMBRA, all other participating systems used external resources (such as Wikipedia and Google n-grams); it is thus fair to assume they had access to at least as much training data as our SCAN model. Consequently, the gap in performance can not solely be attributed to a difference in the size of the training data.

We also observe that IXA and SCAN, given identical class granularity, perform better on subtask 1, while AMBRA and our own SVM-based systems exhibit the opposite trend. The IXA system uses a combination of knowledge sources in order to determine when a piece of news was written, including explicit mentions of temporal expressions within the text, named entities, and linked information to those named entities from Wikipedia. AMBRA on the other hand exploits more shallow stylistic, grammatical and lexical features within the learning-to-rank paradigm. An interesting direction for future work would be to investigate which features are most appropriate for different DTE tasks. Overall, it is encouraging to see that the generic temporal word representations inferred by SCAN lead to competitively performing models on both temporal classification tasks without any explicit tuning.

6.4.5 Discussion

We applied SCAN, a dynamic Bayesian model of sense development to the phenomenon of diachronic word meaning change. Our model learns a coherent set of co-dependent time-specific senses for individual words and their prevalence. Evaluation of the model output showed that the learnt representations reflect (a) different senses of ambiguous words (b) different kinds of meaning change (such as new senses being established), and (c) connotational changes within senses. SCAN departs from previous work in that it models temporal dynamics explicitly. We demonstrated that this feature yields more general semantic representations as indicated by predictive log-likelihood and a variety of extrinsic evaluations. We also experimentally evaluated SCAN on novel sense detection and the SemEval DTE task, where it performed on par with the best published results, without any extensive feature engineering or task specific tuning.

In our experiments we used context as a bag of words. It would be interesting to explore more systematically how different kinds of contexts (e.g., named entities, multiword expressions, verbs vs. nouns) influence the representations the model learns. Furthermore, while SCAN captures the temporal dynamics of word *senses*, it cannot do so for words themselves. Put differently, the model cannot identify whether a new word is used which did not exist before, or that a word ceased to exist after a specific point in time. A model internal way of detecting word (dis)appearance would be desirable, especially since new terms are continuously being introduced thanks to popular culture and various new media sources.

6.5 Summary

This chapter introduced a novel Bayesian model of dynamic sense change, SCAN, which infers a globally coherent representation of gradual meaning development of individual words over time. We presented computational investigations of two phenomena pertaining to the dynamic nature of meaning representations: Firstly, we modeled meaning development 'in the small' by exploring how young children acquire the meaning of concepts and how child-like conceptual representations develop over time to resemble established adult-like representations. Secondly, we investigated meaning change 'in the large', studying the process of diachronic change in word meaning over decades and centuries.

In order to investigate the dynamic nature of child conceptual representations during learning, we presented our model with child-directed language and analyzed the development of featural representations of concepts over time. Our model learns from concept mentions in their linguistic context, where we use the context as an approximation of perceived features. We showed for a broad range of concepts and features that concept representations increase in complexity, that phenomena such as diversification of meaning representation emerge from our model. We can conclude that child-directed language encodes the necessary structure and information that drives the acquisition and development of concept meaning.

In addition, we applied our model to historical data and modeled semantic change of word meaning over time. In contrast to previous models we explicitly capture the smooth and gradual nature of meaning change. We demonstrated the benefit of this modeling decision both qualitatively through learnt time-specific word representations that are intuitively interpretable, and quantitatively in a series of diverse experiments. Our general model, developed without a particular semantic task in mind, performs competitively with related models of word meaning change across evaluations.

Chapters 4 and 5 introduced cognitively plausible Bayesian models of category acquisition together with incremental learning algorithms which approximate the on-line nature of human learning. While in this chapter we proposed a cognitively motivated model for child feature acquisition from natural language text, we used a batch learning algorithm which stores, and repeatedly iterates over, all the training data available. In order to investigate the behavior of our model under more realistic constraints, it would be desirable to incrementalize our Gibbs sampler for the SCAN model as well. However, we leave this for future work.

Another interesting direction for future work concerns the application of our model to monitoring meaning change in contexts beyond the cognitive and diachronic settings studied in this chapter. We could apply our model to different text genres and levels of temporal granularity. For example, we could work with Twitter data, an increasingly popular source for opinion tracking, and use our model to identify short-term changes in word meanings or connotations. Investigating feature acquisition and development from multi-modal data (e.g., comprising visual and linguistic information) could be another interesting continuation of this work.

Chapter 7

Conclusions

This chapter concludes the thesis with a summary of our main findings (Section 7.1), and outlines future research directions (Section 7.2).

Humans constantly form and adapt knowledge about their complex environment. Categories provide an efficient way for storing and using knowledge about the world around us, and are integral to how we perceive and interact with our surroundings. Given their fundamental nature, questions of how categories are acquired, mentally represented, and dynamically adapted have received much attention in prior research. Previous behavioral and computational research has mostly involved a small number of toy stimuli (such as strings of binary numbers) with carefully controlled features. This stands in sharp contrast to the complexity of the environment which categories are supposed to capture. This thesis takes a step towards bridging this gap.

There is ample evidence that the acquisition of language and conceptual knowledge are tightly intertwined problems which mutually guide and boost each other (Chapter 2). Based on this insight we model the acquisition and representation of categories based on natural language input. Specifically, we use corpora (including data sets of transcribed child-directed speech from child-parent interactions) to represent the learning environment from which categories are acquired and from which structured representations of categories are learnt. In our case, concept observations amount to their linguistic mentions in corpora, and concept features are represented by the linguistic context in which concepts occur. Based on these assumptions we developed three novel Bayesian models of categorization-related aspects which we evaluated based on their ability to acquire and represent categories comprising hundreds of concrete natural concepts.

Before summarizing our main findings, it is worth discussing the angle of model evaluation chosen throughout this thesis, and the limitations it entails. We began this thesis noting that humans acquire knowledge with a remarkable efficiency under cognitive constraints: humans learn incrementally, and are subject to memory constraints. In addition, children learn categories 'from scratch'. They (arguably) have no access to prior category knowledge to start with, and cannot process the input they receive in sophisticated ways (e.g., they cannot syntactically interpret language). These observations motivate two research questions:

1. Can we build cognitively motivated computational models that incorporate the above constraints, and efficiently learn representations of high quality?
2. Do humans behave in ways that are consistent with the predictions made by the models, beyond their ability to learn successfully (e.g., in terms of the kinds of categories learnt, or the order in which they emerge)?

The models presented in this thesis were evaluated predominantly with respect to question 1. We evaluated the output of our models against a human-produced gold standard of categories, and human-produced plausibility judgments of the acquired featural representations. We also compared our own models quantitatively against previous models of category and feature learning, and showed that they perform competitively across evaluation tasks.

Our evaluations did not shed light on the question of whether the types of emerging representations and their developmental process predicted by our models are consistent with human behavior: Do children learn all and only the categories our model predicts? Do intermediate category representations resemble those of children in the process of category acquisition? The overarching goal of this thesis was to model the acquisition and development of categories and features on a scale and representational complexity approaching the characteristics of the environment from which humans learn. We are not aware of a behavioral data set on this scale, and creating such a data set would be a major undertaking on its own which we leave for future work.

Taken together, the results presented in this thesis show that natural language input encodes the structure that drives human category learning, and that our Bayesian models are able to distill the relevant information from naturalistic data on a large scale. We believe that our work takes a step towards understanding how humans utilize the complex structure of their environment to construct conceptual knowledge.

7.1 Main Findings

In the following we summarize the central findings of this work.

Category learning. Chapters 4 and 5 investigated the acquisition of a large number of categories of concrete natural objects. We introduced two cognitively motivated Bayesian models of child category acquisition, BayesCat and BCF. Our models are knowledge-lean (they do not assume sophisticated linguistic processing abilities such as parsing), they learn categories 'from scratch' (no explicit category knowledge is instilled in the model in the beginning of the learning process), and they learn in an unsupervised way. We combine our models with a cognitively motivated algorithm for approximate inference (particle filtering), modeling category learning as an incremental process which integrates novel information as the data is observed. We find that our models capture the human category learning process in various aspects. First, analysis of the incremental learning algorithm revealed that it performs well under the time- and memory constraints reminiscent of human learning. Secondly, our models learn plausible categories when compared against a human-created gold standard. Thirdly, our models simulate the incremental human learning process by learning representations that consistently improve over time, and by acquiring representative features for categories together with the categories themselves. A previously proposed graph-based model of incremental category learning was shown to qualitatively and quantitatively fit human category learning less closely than our models. In summary, our results provide further evidence to the claim that humans acquire categories by aggregating information over time and by establishing representations which describe their environment increasingly accurately.

Feature learning. Categories are not learnt in isolation. Chapter 5 computationally investigates the acquisition of conceptual knowledge in a broader context. We introduce BCF, a model which not only explains the acquisition of categories, but also accounts for the emergence of structured featural representations. Our model captures the joint emergence of (1) categories themselves, (2) their feature representations structured into *types* of relevant properties, and (3) the association of feature types with categories. To the best of our knowledge our work is the first to investigate these phenomena jointly using large-scale naturalistic input. Note that we do not assume that our models induce sets of necessary and sufficient features of concepts and categories in the classical sense. Rather, we argue that the learnt representations capture relevant associated information in the spirit of feature norms (McRae et al., 2005) which have been shown to provide a valuable window into mental representations of concepts and cateogries. Experimental results reveal the effectiveness of our model, and the benefit of joint category and feature learning. The structured features acquired by our model are judged more interpretable by humans, compared to feature types induced by a model which learns categories and features in two separate processes, and is cognitively less plausible in the sense that it requires the availability of a hand-crafted set of rules based on substantial linguistic knowledge for feature detection. We also showed that our model captures the joint emergence of categories and their structured features in infants incrementally when exposed to corpora of child-directed language. The results presented in this thesis suggest that cognitive models capture aspects of the acquisition of complex category representations, and lead us to believe that the debate of the emergence and representation of knowledge can be advanced through large-scale computational investigations.

Meaning Development. A common assumption in previous models of knowledge formation is that concepts are represented through a fixed set of features, however, human conceptual representations can adapt to a changing environment. Chapter 6 presented SCAN, a dynamic Bayesian model of meaning change. Our model infers time-specific representations of concept meaning, and accounts for the temporal dynamics underlying their development. We demonstrated the effectiveness of our model on two tasks. We used SCAN to investigate word meaning change over centuries. Results show not only that the inferred time-specific word representations reveal intuitive and temporally relevant aspects of word meaning, but also that our model performs competitively across a range of semantic evaluation tasks when compared with previously developed task-specific systems. In addition, we modeled the development of concept representations in young infants who form a representation of their environment for the first time. We showed that the representations inferred by our model resemble the increasing complexity of concept representations, a development that has been observed in the behavioral studies with children. This thesis presents the first large-scale computational study of concept development in children that we are aware of.

Natural language input. All experiments presented in this thesis are based on naturalistic language input as a representation of the environment from which categories and their representations are learnt. Does language capture the environmental structure that drives this acquisition process? In line with previous findings which suggest that non-linguistic information from the environment is redundantly encoded in language (Riordan and Jones, 2011), our experiments provide evidence in favor of a positive answer to this question. Our results furthermore support the view that language influences category and feature learning: learners use statistical cues from word usage in context to infer information about categories and their representation. Language corpora are available in large quantities and for a variety of genres. We evaluated our models in two settings: on large collections of general (news or encyclopedic) text, and on corpora of transcribed child-directed speech. This allowed us to compare the representations that our models can learn from data of a different quality (speech data is much noisier than news text) and content (encyclopedic data is created with the purpose of describing knowledge, whereas child-directed speech conveys knowledge implicitly). Applying models to different kinds of corpora can help explain the influence of the input on the acquired representations for different groups of learners, and can provide further insight into the cognitive development of children and adults.

Learning at scale. Bayesian models of category acquisition have primarily been tested on small data sets of artificial stimuli (Anderson, 1991; Sanborn et al., 2006). We showed that Bayesian models capture phenomena of category acquisition when the scope of the learning problem scales both in terms of the number of categories and concepts to be acquired, as well as in terms of their complexity. Our models learn categories comprising hundreds of natural concepts from thousands of linguistic stimuli. Our models learned categories of natural concrete concepts, which have rich and structured features.

Bayesian modeling. This thesis used the framework of Bayesian modeling to investigate the acquisition of conceptual knowledge. We introduced three generative models structured around the incremental, joint and dynamic nature of the acquisition of categories and their features from naturalistic data. Bayesian models formalize probabilistic inference on sets of observed data as a way of inductive learning. Evaluation showed that our models acquire categories which match those encoded in a human-created gold standard, as well as rich and structured sets of relevant associated features. At the same time our models capture the dynamic and incremental process of category acquisition. Thus our results provide further evidence for the view of category and feature learning as instances of statistical inductive inference. Taken together, the experiments presented in this thesis lead us to conclude that Bayesian modeling is a fruitful framework for testing hypotheses about category acquisition, structure, and development.

7.2 Limitations and Directions for Future Research

The framework for modeling category acquisition and representation adopted in this thesis involves a number of assumptions and simplifications. We start by pointing out limitations that these assumptions introduce to our models, and discuss possible improvements (Sections 7.2.1–7.2.2). We conclude with highlighting directions for future research (Sections 7.2.3–7.2.5).

7.2.1 Non-parametric Models of Categorization

Non-parametric Bayesian models can adapt their structure to the complexity of the input data. Previous models of category or features learning have used this feature, allowing the model to adaptively increase in complexity if demanded by the structure of the input data (Anderson, 1991; Sanborn et al., 2006; Austerweil and Griffiths, 2013). In contrast, the models developed in this thesis are parametric, the number of categories or the structure of feature representations is determined a priori. Beyond inferring the number of categories, non-parametric extensions of our models could capture other representational aspects of categories more realistically. In Chapter 5 we assumed that every category is represented by the same number of feature types. A non-parametric model would lift this assumption. A non-parametric model of dynamic meaning change (Chapter 6) could *explicitly* capture the emergence and disappearance of featural aspects of concepts in child language acquisition; or the birth and death of word senses in diachronic word meaning change.

7.2.2 Integrating Word and Category Learning

The models presented in this thesis learn from collections of natural language stimuli consisting of a target concept mention and its surrounding context. This input is based on the rather bold assumption that the learner has solved a significant part of the word learning problem: she has successfully mapped each target concept to a word. As discussed in detail in Chapter 2, word learning itself constitutes a big challenge for young infants. Our work remains agnostic about the fact that the meaning of words itself needs to be acquired, and that knowledge about concepts and categories will help tackle the word learning problem. A fully faithful model would consider the problems of word and concept or category learning jointly. Extending our models to account for this joint optimization will be a very interesting avenue for future research.

7.2.3 Representation of the Learning Environment

In this thesis we used natural language input as an approximation of the environment from which categories and their representations are learnt. While we showed that the linguistic environment is a useful *approximation* of the full multimodal input a learner has access to, it is clear that this multimodal environment is not *fully* captured in language. Computational models of word learning have been trained on multimodal input data (albeit on smaller-scale problems; Frank et al. 2009; Yu and Smith 2007). Advantageously, Bayesian models are flexible with respect to the input data they receive, so we expect the application of our models to multimodal data to be a feasible avenue for future work. Applying our models to such data sets would allow to compare the category acquisition process and the acquired representations which emerge from models trained on multimodal input against those emerging from purely linguistic data.

7.2.4 Learning Abstract Categories

Humans not only categorize the physical world around them, but also infer complex representations of abstract categories and concepts such as POLITICAL (e.g., *parliament*, *socialist*), LEGAL (e.g., *law*, *trial*), or FEELINGS (e.g., *mirth* or *embarrassment*). Lacking any physical realization, and hence perceivable properties, it is to be expected that language plays a particularly important role in acquiring the meaning of such abstract concepts (Wiemer-Hastings and Graesser, 2000). Using the models presented in this thesis to learn classes of abstract concepts and their structured representations is an obvious extension. The SCAN model of dynamic meaning change (Chapter 6) could also be used to infer the change of connotations with abstract political concepts and ideas.

7.2.5 Category Acquisition across Languages and Cultures

One advantage of modeling knowledge acquisition from text is its generalizability across languages. Linguistic corpora are available in large quantities for many languages, including corpora of child-directed speech. Since our models are knowledge lean (i.e., they do not require sophisticated linguistic pre-processing tools), they are straightforwardly applicable across languages. Do children acquire concepts and categories in different orders across cultures? Do the same categories emerge at all? Especially in the context of abstract categorization (7.2.4), these questions provide interesting potential for future research.

Appendix A

Derivation of the Gibbs Sampler for Dirichlet-Multinomial Distributions

We first show how to analytically integrate over Multinomial parameters in Dirichlet-Multinomial models. Afterwards, we derive the full-conditional update equations for collapsed Gibbs sampling. We derive these equations for a model reminiscent of Naive Bayes, which is the simplest and most similar model to the models introduced in this thesis.

Naive Bayes is a model for classifying observations into a fixed and discrete set of classes $k = 1...K$. It assumes observations represented as sets of independent features (e.g., documents as sets of terms) $d = [w_1, ..., w_N]$,[1] and assigns one class label z^d to each document. Naive Bayes assigns class labels to documents based on (a) the a priori probability of a label z, and (b) the probability of the observed terms w given z. In terms of the generative story, we first draw a label z from a Multinomial distribution over labels $Mult(\theta)$. Afterwards we draw iid. terms w_i from a class-specific Multinomial distribution over terms $Mult(\phi_z)$. All Multinomial parameters are drawn from Dirichlet priors:

$$\begin{aligned} \theta &\sim Dir(\alpha) & \qquad z &\sim Mult(\theta) \\ \phi_k &\sim Dir(\beta) & \qquad w_i &\sim Mult(\phi_z). \end{aligned}$$

[1] Analogously, Naive Bayes can model observations of objects, or concepts, as sets of features. Throughout this derivation we will use the document-term example and terminology.

We start from the joint distribution,

$$\begin{aligned} p(\mathbf{d},\mathbf{z},\theta,\{\phi\}_1^Z;\alpha,\beta) = \\ Dir(\theta|\alpha)\prod_z Dir(\phi_z|\beta)\prod_d Mult(z^d|\theta)\prod_d\prod_i Mult(w_i^d|\phi_{z^d}) \end{aligned} \tag{A.1}$$

We will first show how we analytically compute the integrals (getting rid of any explicit $\int$ in our formula).[2] We start by integrating over θ and ϕ and re-grouping the factors in equation (A.1) according to their dependencies on these parameters:

$$\begin{aligned} p(\mathbf{d},\mathbf{z};\alpha,\beta) &= \int_\theta\int_\phi p(\mathbf{d},\mathbf{z},\theta,\{\phi\}_1^Z;\alpha,\beta)d\theta d\phi \\ &= \int_\theta p(\theta|\alpha)\prod_d p(z^d|\theta)d\theta \;\times\; \int_\phi \prod_z p(\phi_z|\beta)\prod_d\prod_i p(w_i^d|\phi_{z^d})d\phi \qquad (A.2) \\ &= \int_\theta p(\theta|\alpha)\prod_d p(z^d|\theta)d\theta \;\times\; \prod_z\int_{\phi_z} p(\phi_z|\beta)\prod_d\prod_i p(w_i^d|\phi_{z^d})d\phi_z. \end{aligned}$$

This shows that θ as well as all ϕ_z are independent under the model. We go through the analytic integration for θ, the parameters of the multinomial probability distribution over classes. The derivation for each ϕ_z (the distribution over terms v for a particular class z) is identical.

$$\int_\theta p(\theta|\alpha)\prod_d p(z^d|\theta)d\theta$$

$$= \int_\theta Dir(\theta|\alpha)\prod_d Mult(z^d|\theta)d\theta \tag{A.3}$$

$$= \int_\theta \frac{\Gamma(\sum_z\alpha)}{\prod_z\Gamma(\alpha)}\prod_z\theta_z^{\alpha-1+n_z}d\theta \tag{A.4}$$

$$= \frac{\Gamma(\sum_z\alpha)}{\prod_z\Gamma(\alpha)}\frac{\prod_z\Gamma(n_z+\alpha)}{\Gamma(\sum_z n_z+\alpha)}\int_\theta\frac{\Gamma(\sum_z n_z+\alpha)}{\prod_z\Gamma(n_z+\alpha)}\prod_z\theta_z^{\alpha-1+n_z}d\theta \tag{A.5}$$

$$= \frac{\Gamma(\sum_z\alpha)}{\prod_z\Gamma(\alpha)}\frac{\prod_z\Gamma(n_z+\alpha)}{\Gamma(\sum_z n_z+\alpha)} \tag{A.6}$$

$$\propto \frac{\prod_z\Gamma(n_z+\alpha)}{\Gamma(\sum_z n_z+\alpha)}, \tag{A.7}$$

where we first, in eqn (A.3)–(A.4),write out the Dirichlet-Multinomial (from eqn 3.10, 3.11). In (A.5) we add factors (before and after the integral) which cancel out, i.e., do not change the equation, but allow us to evaluate the terms inside the integral to 1 so that it can be dropped in (A.6). We finally drop all constants that do not depend

[2] This derivation is based on Carpenter (2010), who provides more detailed explanations.

on z. We have now eliminated the integral, and represent θ implicitly as counts of class-assignments (n_z) to the data.

Following the same procedure we can derive for the set of $\{\phi_z\}$

$$\prod_z \int_{\phi_z} p(\phi_z|\beta) \prod_d \prod_i p(w_i^d|\phi_{z^d}) d\phi_z \quad \propto \quad \prod_z \frac{\prod_v \Gamma(n_v^z + \beta)}{\Gamma(\sum_v n_v^z + \beta)}, \tag{A.8}$$

where n_v^z refers to the count of term w occurring with an document labeled with class z.

In Gibbs sampling, we are interested in sampling each individual z_i from its full conditional distribution. We will now derive the full conditional distribution over all possible values z_j, and show that it has a mathematically simple form with an intuitive explanation. We resample the class z^j for document j given the class assignments to all other documents z^{-j}, the data d and hyperparameters. As shown in equation (3.29) this distribution is proportional to the joint distribution derived above, so using (A.7) and (A.8) we can write:

$$p(z^j|z^{-j}, d, \alpha, \beta) \quad \propto \quad \frac{\prod_z \Gamma(n_z + \alpha)}{\Gamma(\sum_z n_z + \alpha)} \times \prod_z \frac{\prod_v \Gamma(n_v^z + \beta)}{\Gamma(\sum_v n_v^z + \beta)}. \tag{A.9}$$

Since the probability $p(z^j)$ is conditioned on the current class assignments to all documents except j, the counts (n_r and n_v^r) regarding any class $r \neq z^j$ are not affected. We split the terms which depend on z^j (the value of the class assigned to document j; terms 2 and 4 in A.10) from those which do not (i.e., concerning all classes $r \neq z^j$; terms 1 and 3 in A.10), and update only the counts regarding z^j: n_{z^j} is incremented by 1 (because it is assigned to one additional document j) and the count of observing any term v with class z^j, ($n_v^{z^j}$), is incremented by c_j^v, the number of times term v occurs in document j,

$$\begin{aligned} &\frac{\prod_{r \neq z} \Gamma(n_r^{-j} + \alpha)}{\Gamma(1 + \sum_r n_r^{-j} + \alpha)} \times \Gamma(n_{z^j}^{-j} + \alpha + 1) \times \\ &\qquad \prod_{r \neq z} \frac{\prod_v \Gamma(n_v^{r,-j} + \beta)}{\Gamma(\sum_v n_v^{r,-j} + \beta)} \times \frac{\prod_v \Gamma(n_v^{z^j,-j} + \beta + c_j^v)}{\Gamma(c_j^v + \sum_v n_v^{z^j,-j} + \beta)}. \end{aligned} \tag{A.10}$$

By definition, $\Gamma(x+q) = \Gamma(x) \prod_{i=1}^{q} (x+i)$. We use this fact to pull apart terms 2 and 4 in (A.10) accordingly.[3] This slightly complicates the equation but will eventually lead

[3] Or as a special case $\Gamma(x+1) = x\Gamma(x)$. We need the general form for the class-term counts because each count can be greater than one.

to a significant simplification:

$$
\begin{aligned}
&= \frac{\prod_{r\neq z}\Gamma(n_r^{-j}+\alpha)}{\Gamma(1+\sum_r n_r^{-j}+\alpha)} \times \Gamma(n_{z^j}^{-j}+\alpha)(n_{z^j}^{-j}+\alpha)\times \\
&\prod_{r\neq z}\frac{\prod_v \Gamma(n_v^{r,-j}+\beta)}{\Gamma(\sum_v n_v^{r,-j}+\beta)} \times \frac{\prod_v \left(\Gamma(n_v^{z^j,-j}+\beta)\prod_{i=1}^{c_v^j}(n_v^{z^j,-j}+\beta+i)\right)}{\Gamma(\sum_v n_v^{z^j,-j}+\beta)\prod_{i=1}^{c^j}(\sum_v n_v^{z^j,-j}+\beta+i)}.
\end{aligned}
\tag{A.11}
$$

Here c_v^j refers to the count of term v in document j, like above, and c^j refers to the total number of terms in j. Next, we conflate all $\Gamma()$ functions over $r \neq z^j$ with those over z^j,

$$
\begin{aligned}
&= \frac{\prod_r \Gamma(n_r^{-j}+\alpha)}{\Gamma(1+\sum_r n_r^{-j}+\alpha)} \times (n_{z^j}^{-j}+\alpha)\times \\
&\prod_r \frac{\prod_v \Gamma(n_v^{r,-j}+\beta)}{\Gamma(\sum_v n_v^{r,-j}+\beta)} \times \frac{\prod_v \prod_{i=1}^{c_v^j}(n_v^{z^j,-j}+\beta+i)}{\prod_{i=1}^{c^j}(\sum_v n_v^{z^j,-j}+\beta+i)}.
\end{aligned}
\tag{A.12}
$$

All $\Gamma()$ components are now constant with respect to any particular value for z^j, such that they can be dropped and we finally obtain:

$$
p(z^j|z^{-j},d,\alpha,\beta) \quad \propto \quad (n_{z^j}^{-j}+\alpha)\times \frac{\prod_v \prod_{i=1}^{c_v^j}(n_v^{z^j,-j}+\beta+i)}{\prod_{i=1}^{c^j}(\sum_v n_v^{z^j,-j}+\beta+i)}. \tag{A.13}
$$

This is a very intuitive result: the probability that document j belongs to class z^j is proportional to the number of times class z^j is assigned to any other document (first term), times the number of times each individual term of document j was observed under class z (second term). Term observation-wise increments i increase the likelihood of terms under repeated observation: the second observation of a term with class z^j is intuitively more likely than observing it for the first time. All counts are smoothed by the Dirichlet parameters, α and β, respectively.

Appendix B

Instructions for Mechanical Turk Experiments

We provide the instructions given in the Mechanical Turk experiments reported in Sections 5.3.3.

B.1 Feature Type Intrusion Task

Please Note

- You have to be a **native speaker of English** to take part in this study.
- In order to receive payment, you have to label and rate all feature sets, **all fields are required.**
- You are welcome to **complete as many hits as you like.**
- Please do not forget to **accept the HIT** before you start working on it.

Informed Consent

This is a linguistic experiment performed at the University of Edinburgh. If you have any questions about this study, feel free to contact Lea Frermann (l.frermann at ed.ac.uk). Participation in this research is voluntary. You have the right to withdraw from the experiment at any time. The collected data will be used for research purposes only. Personal data will be kept confidential and will not be shared with third parties.

Personal Details Questionnaire

Please fill in the Personal Details questionnaire correctly, as otherwise you will not receive payment.

1. Age:
2. Gender:
3. Please specify the country where you have learned your first language:

Instructions

Categories such as ANIMAL or FURNITURE are represented by example concepts (e.g., cat, dog, bed, table) and can be described in terms of their features or attributes. For example ANIMALS can be found in locations such as {forests, gardens, trees} or have visual properties, such as {fur, legs, ears, feathers}. FURNITURE, on the other hand, is typically found in locations such as {stores, living rooms, kitchens} and has external properties such as {seats, legs, patterns}.

Your Task

In this experiment, you will be presented with concepts (e.g., cat, dog) exemplifying a category (e.g. ANIMAL) and different feature collections describing this category (e.g., {fur, legs, ears, feathers}, {forests, gardens, trees}). **One** of the feature collections is **not applicable** to this category. Your task is to **detect the feature collection which does not belong to the category**.

Please do not forget to accept the HIT before you start working on it.

B.2 Word Intrusion Task

Please Note

- You have to be a **native speaker of English** to take part in this study.
- In order to receive payment, you have to label and rate all feature sets, **all fields are required.**
- You are welcome to **complete as many hits as you like.**
- Please do not forget to **accept the HIT** before you start working on it.

Informed Consent

This is a linguistic experiment performed at the University of Edinburgh. If you have any questions about this study, feel free to contact Lea Frermann (l.frermann at ed.ac.uk). Participation in this research is voluntary. You have the right to withdraw from the experiment at any time. The collected data will be used for research purposes only. Personal data will be kept confidential and will not be shared with third parties.

Personal Details Questionnaire

Please fill in the Personal Details questionnaire correctly, as otherwise you will not receive payment.

1. Age:

2. Gender:

3. Please specify the country where you have learned your first language:

Instructions

In this experiment, you will be presented with groups of words which refer to **one common topic**. However, each group contains **one word which does not belong to this topic**. Your task is to detect this “intruder word”. Please select the intruder based on the **meaning** of the word, and not its part-of-speech or spelling. If you think multiple words do not belong to the group please use your best judgment for selecting the best candidate. You must **select one intruder word for every group** of words.

Examples

{apple banana car orange pear}	The word “car” is the intruder word since it does not belong to the general topic of fruit.
{keyboard screen yellow write laptop}	The word “yellow” is the intruder word since it does not belong to the general topic of office equipment/work.

Please do not forget to accept the HIT before you start working on it.

Appendix C

Additional Material on Experiment 5

C.1 Set of Target Concepts

The table below lists the set of target words used in Experiment 5, the study on dynamic feature development in language acquisition (Section 6.3). Most concepts are basic-level categories taken from the McRae concept set based on frequency of occurrence in the training corpus. Exceptions are marked with an (*) and comprise superordinate-level categories, one abstract noun, adjectives and verbs.

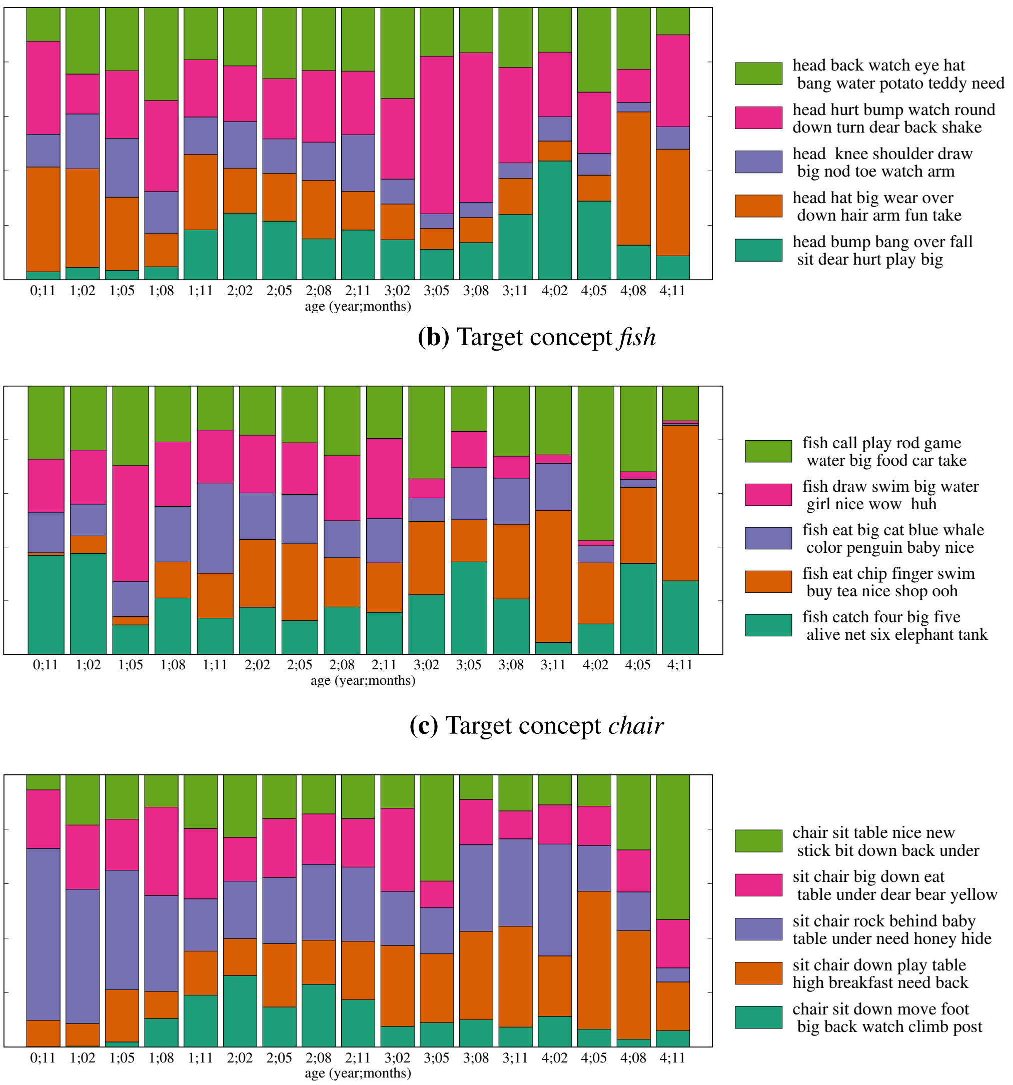
ID	concept	ID	concept	ID	concept
1	animal*	11	color*	21	head
2	apple	12	dog	22	horse
3	bag	13	door	23	house
4	bed	14	eat*	24	nose
5	bedroom	15	fish	25	orange
6	blue*	16	food*	26	play*
7	box	17	green*	27	red*
8	car	18	hair	28	table
9	cat	19	hand	29	toy*
10	chair	20	hat	30	train

C.2 Additional Model Output

We provide example SCAN representations in addition to the output discussed in Section 6.3.2 Time-specific meaning representations are visualized as a bar capturing the relative prevalence ($p(k|t) = \phi_k^t$) of different feature types (color-coded). One such visualization is displayed for each temporal interval, illustrating the development of feature type prevalence over time. Each interval covers $\Delta t = 3$ months, and is labeled with the start date, i.e., age of the child. Each feature type is illustrated to the right of the plot as the ten words w most highly associated with the feature type, marginalizing over the time-specific representations $\left(p(w|k) = \sum_t \psi_w^{t,k}\right)$.

The Full corpus. Example output of SCAN trained on the conflated input to 21 children for the concepts *head*, *fish*, *chair*, *dog* and *bag*.

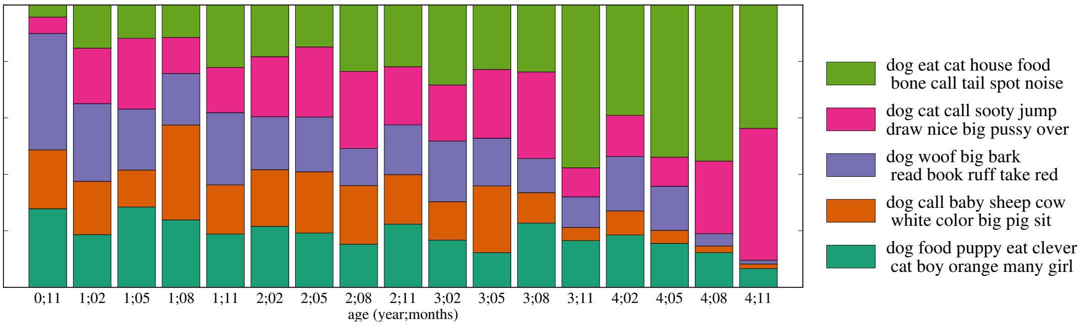
(a) Target concept *head*



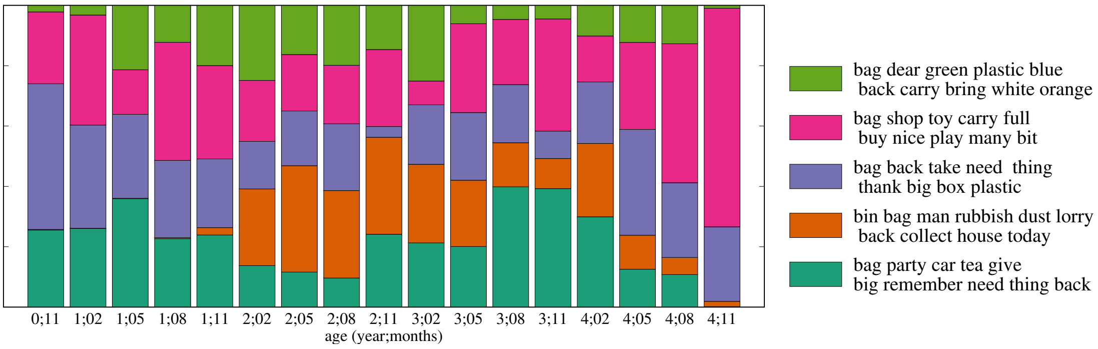
(b) Target concept *fish*

(c) Target concept *chair*

(d) Target concept *head*

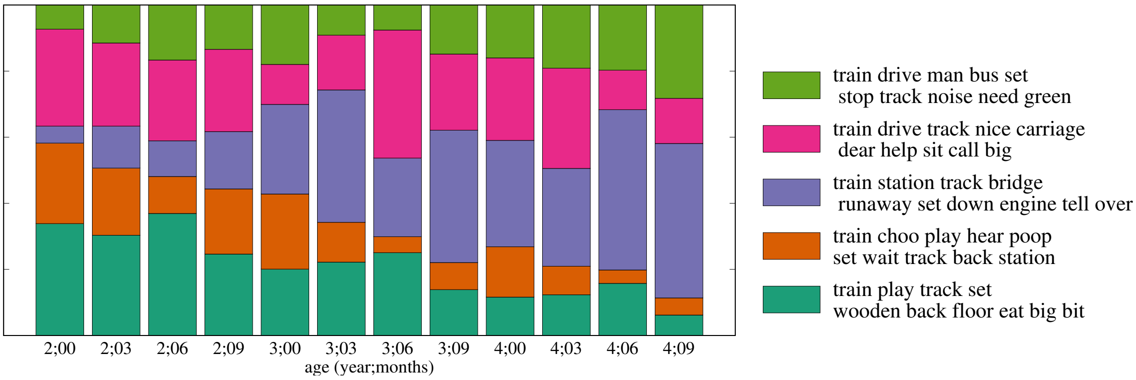


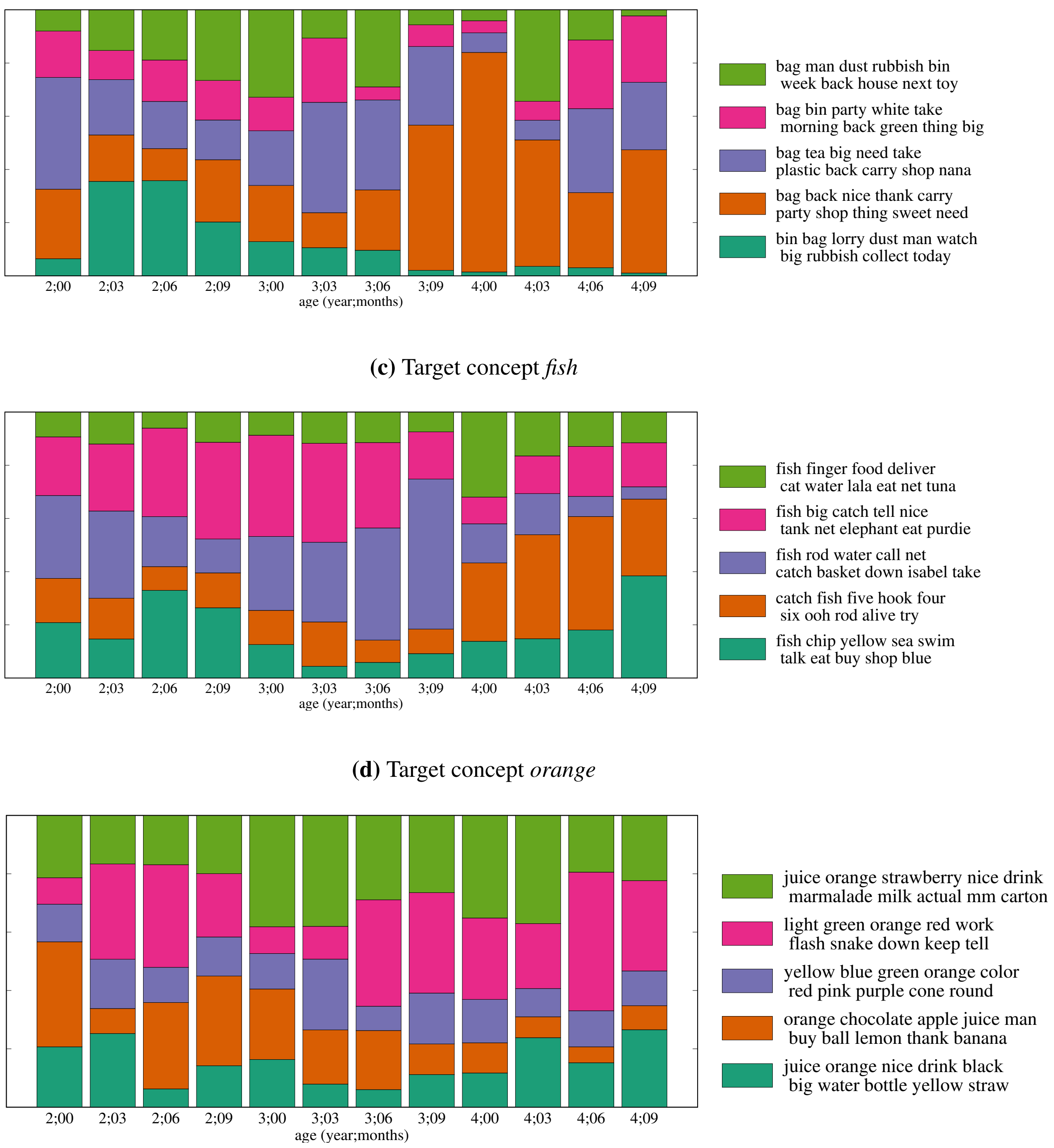
(e) Target concept *bag*



The Thomas corpus. Example output of SCAN trained on the Thomas corpus for the concepts *train*, *bag*, *fish* and *orange*.

(a) Target concept *train*





(b) Target concept *bag*

(c) Target concept *fish*

(d) Target concept *orange*

Bibliography

Ahn, W.-K. (1998). Why are different features central for natural kinds and artifacts?: the role of causal status in determining feature centrality. *Cognition*, 69:135.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177.

Aitchison, J. (2001). *Language Change: Progress Or Decay?* Cambridge Approaches to Linguistics. Cambridge University Press.

Akhtar, N. and Tomasello, M. (2000). The social nature of words and word learning. In Golinkoff, R. M., Hirsh-Pasek, K., Bloom, L., Smith, L., Woodward, A., Akhtar, N., Tomasello, M., and Hollich, G., editors, *Becoming a word learner: A debate on lexical acquisition*, pages 115–135.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Ashby, F. and Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39:629–654.

Austerweil, J. L. and Griffiths, T. L. (2009). Analyzing human feature learning as nonparametric Bayesian inference. In *Advances in Neural Information Processing Systems (NIPS)*, pages 97–104.

Austerweil, J. L. and Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive Psychology*, 63(4):173–209.

Austerweil, J. L. and Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, 120(4):817–851.

Baillargeon, R. (1987). Young infants' reasoning about the physical and spatial properties of a hidden object. *Cognitive Development*, 2(3):179–200.

Baroni, M. (2010). Detailed description of the Strudel algorithm. Technical report.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M., Lenci, A., and Onnis, L. (2007). ISA meets Lara: an incremental word space model for cognitively plausible simulations of semantic learning. In *Proceed-*

ings of the Workshop on Cognitive Aspects of Computational Language Acquisition, pages 49–56. Association for Computational Linguistics.

Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11:629–654.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In Neisser, U., editor, *Concepts and conceptual development: Ecological and intellectual factors in categorization*, pages 101–140. Cambridge University Press, Cambridge.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–609.

Biemann, C. (2006). Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In *Proceedings of TextGraphs: the 1st Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly Media, Inc., 1st edition.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Blei, D. M. and Lafferty, J. D. (2006a). Correlated Topic Models. In *Advances in Neural Information Processing Systems*, pages 147–154. Morgan Kaufmann Publishers Inc., Vancouver, BC, Canada.

Blei, D. M. and Lafferty, J. D. (2006b). Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, Pittsburgh, PA, USA.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bomba, P. C. and Siqueland, E. R. (1983). The nature and structure of infant form categories. *Journal of Experimental Child Psychology*, 35:294–328.

Bornstein, M. H. and Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development*, 81(3):884–897.

Borovsky, A. and Elman, J. (2006). Language input and semantic categories: a relation between cognition and early word learning. *Journal of Child Leanguage*, 33:759–790.

Börschinger, B. and Johnson, M. (2011). A particle filter algorithm for bayesian word segmentation. In *Proceedings of the Australasian Language Technology Association workshop*, pages 10–18.

Börschinger, B. and Johnson, M. (2012). Using rejuvenation to improve particle filtering for Bayesian word segmentation. In *ACL (2)*, pages 85–89. The Association for Computer Linguistics.

Braine, M. D. S. (1987). What is learned in acquiring word classes – a step toward an acquisition theory. In MacWhinney, B., editor, *Mechanisms of language acquisition*, chapter 3, pages 65–87. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Brody, S. and Lapata, M. (2009). Bayesian Word Sense Induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 103–111.

Brown, R. (1958). *Words and things*. The Free Press.

Brown, R. W. (1957). Linguistic determinism and the parts of speech. *Journal of Abnormal and Social Psychology*, 55(1):1–5.

Brown, S. D. and Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58:49–67.

Callanan, M. A. (1990). Parents' descriptions of objects: Potential data for children's inferences about category principles. *Cognitive Development*, 5(1):101 – 122.

Canini, K. (2011). *Nonparametric Hierarchical Bayesian Models of Categorization*. PhD thesis, EECS Department, University of California, Berkeley.

Canini, K. R., Shi, L., and Griffiths, T. L. (2009). Online inference of topics with latent Dirichlet allocation. *Journal of Machine Learning Research - Proceedings Track*, 5.

Carpenter, B. (2010). Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling. Technical report, LingPipe.

Chambers, N. (2012). Labeling Documents with Timestamps: Learning from their Time Expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 98–106, Jeju Island, Korea.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc.

Chapman, K. L., Leonard, L. B., and Mervis, C. B. (1986). The effect of feedback on young children's inappropriate word usage. *Journal of Child Language*, 13:101–117.

Chater, N., Oaksford, M., Hahn, U., and Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):811–823.

Cochran, W. G. (1977). *Sampling Techniques, 3rd Edition*. John Wiley.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Colunga and Sims (2011). Early talkers and late talkers know nouns that license different word learning biases. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 2550–2555.

Colunga, E. and Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, 12(2):347–382.

Connel, L. and Ramscar, M. (2001). Using distributional measures to model typicality in categorization. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, pages 226–231.

Cook, P., Lau, J. H., McCarthy, D., and Baldwin, T. (2014). Novel Word-sense Identification. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland.

Cook, P. and Stevenson, S. (2010). Automatically Identifying Changes in the Semantic Orientation of Words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 28–34, Valletta, Malta.

Corter, J. E. and Gluck, M. A. (1992). Explaining basic categories - feature predictability and information. *Psychological Bulletin*, 111(2):291–303.

Cree, G. S., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, 23(3):371–414.

Csibra, G. and Gergely, G. (2006). *Social learning and social cognition: The case for pedagogy*, pages 249 – 274. Oxford University Press, Oxford.

Davies, M. (2010). The Corpus of Historical American English: 400 million words, 1810-2009. Available online at `http://corpus.byu.edu/coha/`.

Daw, N. D. and Courville, A. (2007). The pigeon as particle filter. In *Advances in Neural Information Processing Systems*, volume 20, pages 369–376. MIT Press.

Demuth, K., Culbertson, J., and Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, 49(2):137–174.

Devereux, B., Pilkington, N., Poibeau, T., and Korhonen, A. (2009). Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. *Research on Language & Computation*, 7(2-4):137–170.

Diaz, M. and Ross, B. H. (2006). Sorting out categories: Incremental learning of category structure. *Psychonomic Bulletin and Review*, 13(2):251–256.

Diller, H.-J., de Smet, H., and Tyrkkö, J. (2011). A European database of descriptors of english electronic texts. *The European English messenger*, 19(2):29–35.

Dominey, P. F. and Dodane, C. (2004). Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics*, 17(2âĂŞ3):121 – 145.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

Doucet, A., de Freitas, N., Murphy, K., and Russell, S. (2000a). Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, pages 176–183. Morgan Kaufmann Publishers Inc.

Doucet, A., Godsill, S., and Andrieu, C. (2000b). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208.

Doucet, A. and Johansen, A. M. (2008). A Tutorial on Particle Filtering and Smoothing: Fifteen years Later. Technical report.

Evans, J. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Essays in cognitive psychology. Psychology Press.

Fahlman, S. E. and Lebiere, C. (1990). The cascade-correlation learning architecture. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 2*, pages 524–532. Morgan Kaufmann Publishers Inc.

Farah, M. J. and McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4):339–357.

Fazly, A., Alishahi, A., and Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063.

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14(1):11–21.

Fei-Fei, L., Fergus, R., and Perona, P. (2003). A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the 8th International Conference on Computer Vision*, volume 2, pages 1134–1141.

Fellbaum, C. (1998a). *WordNet: An Electronic Lexical Database*. Bradford Books.

Fellbaum, C. (1998b). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, MA, USA.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. Wiley, New York.

Fountain, T. (2013). *Modelling the Acquisition of Natural Language Categories*. PhD thesis, ILCC, School of Informatics, University of Edinburgh.

Fountain, T. and Lapata, M. (2010). Meaning representation in natural language categorization. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1916–1921.

Fountain, T. and Lapata, M. (2011). Incremental models of natural language category acquisition. In *Proceedings of the 33nd Annual Conference of the Cognitive Science Society*, pages 255–260.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 457–464.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20:579–575.

Frank, M. C., Tenenbaum, J. B., and Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning, and Development*, 9:1–24.

Frazier, L. and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Frermann, L. and Lapata, M. (2014). Incremental Bayesian Learning of Semantic Categories. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 249–258.

Frermann, L. and Lapata, M. (2015a). A Bayesian Model for Joint Learning of Categories and their Features. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1576–1586. Association for Computational Linguistics.

Frermann, L. and Lapata, M. (2015b). Incremental Bayesian Category Learning from Natural Language. *Cognitive Science*, pages 1–49. early view.

Frermann, L. and Lapata, M. (2016). A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gelman, S. (1988). The development of induction within natural kind and artifact categories. *Cognitive psychology*, 20(1):65–95.

Gelman, S. and Coley, J. (1990). The Importance of Knowing a Dodo Is a Bird: Categories and Inferences in 2-Year-Old Children. *Developmental Psychology*, 26(5):796–804.

Gelman, S. and Keil, F. (1998). *Beyond labeling: the role of maternal input in the acquisition of richly structured categories*. Number no. 253 in Monographs of the Society for Research in Child Development. University of Chicago Press.

Gelman, S. A. and Markman, E. M. (1985). Implicit contrast in adjectives vs. nouns: implications for word-learning in preschoolers. *Journal of Child Language*, 12:125–143.

Gelman, S. A. and Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23(3):183–209.

Gelman, S. A. and Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, 58:1532–1541.

Gelman, S. A. and O'Reilly, A. W. (1988). Children's Inductive Inferences within Superordinate Categories: The Role of Language and Category Structure. *Child Development*, 59(4):876–887.

Gelman, S. A., Wilcox, S. A., and Clark, E. V. (1989). Conceptual and lexical hierarchies in young children. *Cognitive Development*, 4(4):309 – 326.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57(6):1317–1339.

Gilks, W. R. and Berzuini, C. (2001). Following a moving target-monte carlo inference for dynamic bayesian models. *Journal of the Royal Statistical Society Series B*, 63(1):127–146.

Gogate, L. J., Bahrick, L. E., and Watson, J. D. (2000). A Study of Multimodal Motherese: The Role of Temporal Synchrony between Verbal Labels and Gestures. *Child Development*, 71(4):878–894.

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In *Perceptual organization in vision: Behavioral and neural perspectives*, pages 233–278.

Goldstone, R. L., Gerganov, E., L, D., Roberts, M. E., and Goldstone, R. L. (2008). Learning to see and conceive. In *The new cognitive sciences (Part of the Vienna Series in Theoretical Biology)*, pages 163–188, Cambridge, MA. MIT Press.

Goldstone, R. L., Lippa, Y., and Shiffrin, R. M. (2001). Altering Object Representations through Category Learning. *Cognition*, 78:27–43.

Goldstone, R. L. and Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology*, 130:116–139.

Goodman, N., Tenenbaum, J., Feldman, J., and Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154.

Gopnik, A. and Meltzoff, A. (1987). The Development of Categorization in the Second Year and Its Relation to Other Cognitive and Linguistic Developments. *Child Development*, 58(6).

Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing*, 140(2):107–113.

Goswami, U. (2014). *Cognition In Children*. Developmental Psychology: A Modular Course. Taylor & Francis.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007a). Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 323–328.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., and Navarro, D. J. (2008). Categorization as non-parametric bayesian density estimation. In *The Probabilistic Mind: Prospects for bayesian Cognitive Science*, pages 3003–350. Oxford University Press, Oxford, UK.

Griffiths, T. L. and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17:767–773.

Griffiths, T. L., Tenenbaum, J. B., and Steyvers, M. (2007b). Topics in semantic representation. *Psychological Review*, 114:2007.

Groenewald, P. C. N. and Mokgatlhe, L. (2005). Bayesian Computation for Logistic Regression. *Computational Statistics & Data Analysis*, 48(4):857–868.

Gulordava, K. and Baroni, M. (2011). A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, Scotland.

Hall, G. D., R., W. S., and M., H. W. (1993). How Two- and Four-Year-Old Children Interpret Adjectives and Count Nouns. *Child Development*, 64(2):1651–1664.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo methods*. Monographs on statistics and applied probability. Chapman and Hall, London, New York.

Hansen, M. B. and Markman, E. M. (2009). Children's Use of Mutual Exclusivity to Learn Labels for Parts of Objects. *Developmental Psychology*, 45(2):592–596.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(23):146–162.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Heit, E. and Barsalou, L. W. (1996). The instantiation principle in natural categories. *Memory*, 4(4):413–451.

Heit, E. and Rubinstein, J. (1994). Similarity and Property Effects in Inductive Reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20(2):411–422.

Hol, J. D., Schön, T. B., and Gustafsson, F. (2006). On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop*.

Hollich, G., Golinkoff, R. M., and Hirsh-Pasek, K. (2007). Young children associate novel words with complex objects rather than salient parts. *Developmental Psychology*, 43(5):1051–1061.

Huang, Y. and Rao, R. P. (2014). Neurons as monte carlo samplers: Bayesian inference and learning in spiking networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 1943–1951. Curran Associates, Inc.

Humphreys, G. W. and Forde, E. M. E. (2001). Hierarchies, similarity, and interactivity in object recognition: "category-specific" neuropsychological deficits. *Behavioral and Brain Sciences*, 24(3):453–476.

Inhelder, B. and Piaget, J. (1964). *The early growth of logic in the child: classification and seriation*. Routledge and Kegan Paul.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jern, A. and Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, 66:85–125.

Jones, S., Smith, L., and Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child development*, 62(3):499–516.

Kachergis, G., Yu, C., and Shiffrin, R. M. (2014). Developing semantic knowledge through cross-situational word learning. In Bello, P., Guarini, M., McShane, M., and Scassellati, B., editors, *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Keil, F. C. (1987). Conceptual Development and Category Structure. In Neisser, U., editor, *Concepts and conceptual development: Ecological and intellectual factors in categorization*, pages 175–200. Cambridge University Press, Cambridge.

Keil, F. C. (1989). *Concepts, Kinds, and Cognitive Development*. MIT Press.

Kelly, C., Devereux, B., and Korhonen, A. (2014). Automatic extraction of property norm-like data from large text corpora. *Cognitive Science*, 38(4):638–682.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2003). Learning domain structures. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 720–725. Erlbaum.

Kemp, C., Shafto, P., and Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, 64:35–75.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological review*, 99(1):22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, 5:3–36.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Geneva, Switzerland.

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

Landau, B., Smith, L., and Jones, S. (1998). Object perception and object naming in early development. *Trends in Cognitive Science*, 27:19–24.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Lang, J. and Lapata, M. (2011). Unsupervised semantic role induction with graph partitioning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1320–1331.

Lau, H. J., Cook, P., McCarthy, D., Gella, S., and Baldwin, T. (2014). Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 259–270, Baltimore, MD, USA.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word Sense Induction for Novel Sense Detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France.

Lee, M. D. and Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, 9(1):43–58.

Levy, R. P., Reali, F., and Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 937–944. Morgan Kaufmann Publishers Inc.

Lieven, E., Salomo, D., and Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*.

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in mcmc. *Methods in Ecology and Evolution*, 3(1):112–115.

Liu, J. S. and Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044.

Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological review*, 111(2):309–332.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28:203–208.

Lupyan, G., Rakison, D. H., and McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, 18(12):1077–1083.

Macario, J. (1991). Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development*, 6(1):17–46.

MacKay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum Associates, Hillsdale, NJ, USA, third edition edition.

Malt, B. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*, 29(2):85 – 148.

Malt, B. C. and Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, 23(2):250 – 269.

Markman, E. M. (1987). How children constrain the possible meanings of words. In Neisser, U., editor, *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, pages 255–287. Cambridge University Press, Cambridge, GB.

Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In Gelman, S. A. and Byrnes, J. P., editors, *Perspectives on language and thought*, pages 72–106. Cambridge University Press. Cambridge Books Online.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, 92:199 – 227.

Markman, E. M. and Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, 16(1):1 – 27.

Markman, E. M. and Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2):121 – 157.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Co., Inc., New York, NY, USA.

McCloskey, M. and Clucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, 11:1–37.

McMahon, A. M. (1994). *Understanding Language Change*. Cambridge University Press.

McRae, K. and Cree, G. S. (2002). Factors underlying category-specific semantic impairments. In Forde, E. M. E. and Humphreys, G., editors, *Category-specificity in mind and brain*, pages 211–248. Psychology Press.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods*, 37(4):547–59.

Meadows, S. (2006). *The Child as Thinker: the Development and Acquisition of Cognition in Childhood (2nd ed).* Routledge.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In Neisser, U., editor, *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, pages 201–233. Cambridge University Press, Cambridge, GB.

Mihalcea, R. and Nastase, V. (2012). Word Epoch Disambiguation: Finding How Words Change over Time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 259–263, Jeju Island, Korea.

Mimno, D., Wallach, H., and McCallum, A. (2008). Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. In *NIPS Workshop on Analyzing Graphs*, Vancouver, Canada.

Mitra, S., Mitra, R., Maity, S. K., Riedl, M., Biemann, C., Goyal, P., and Mukherjee, A. (2015). An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21:773–798.

Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., and Goyal, P. (2014). That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1020–1029, Baltimore, MD, USA.

Murphy, G. L. (2002). *The Big Book of Concepts.* The MIT Press, Cambridge, MA, USA.

Murphy, G. L., Chen, S. Y., and Ross, B. H. (2012). Reasoning with uncertain categories. *Thinking & Reasoning*, 18(1):81–117.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* The MIT Press, Cambridge, MA.

Navarro, D. J., Perfors, A., and Vong, W. K. (2013). Learning time-varying categories. *Memory & Cognition*, 41(6):917–927.

Neisser, U. (1987). Introduction: Ecological and intellectual factors in categorization. In Neisser, U., editor, *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, pages 1–11. Cambridge University Press, Cambridge, GB.

Norman, G. R., Brooks, L. R., Coblentz, C. L., and Babcook, C. J. (1992). The correlation of feature identification and category judgments in diagnostic radiology. *Memory & Cognition*, 20(4):344–355.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 115:39–57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:700–708.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In Healy, A. F., Josslyn, S. M., and Shiffrin, R. M., editors, *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes*, volume 1, pages 149–167. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Ó Séaghdha, D. (2010). Latent Variable Models of Selectional Preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden.

Paciorek, C. (2009). Understanding intrinsic Gaussian Markov random field spatial models, including intrinsic conditional autoregressive models. Technical vignette. `http://www.stat.berkeley.edu/~paciorek/research/techVignettes/techVignette5.pdf`.

Perfors, A., Kemp, C., and Tenenbaum, J. B. (2005). Modeling the acquisition of domain structure and feature understanding. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1720–1725.

Perfors, A. and Tenenbaum, J. B. (2009). Learning to learn categories. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 136–141.

Pevtzow, R. and Goldstone, R. L. (1994). Categorization and the parsing of objects. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, pages 712–722. Lawrence Erlbaum Associates.

Popescu, O. and Strapparava, C. (2013). Behind the Times: Detecting Epoch Changes using Large Corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 347–355, Nagoya, Japan.

Popescu, O. and Strapparava, C. (2015). SemEval 2015, Task 7: Diachronic Text Evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 869–877, Denver, CO, USA.

Posner, M. I. and Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 21:367–379.

Quinn, P. C. and Eimas, P. D. (1996). Perceptual cues that permit categorical differentiation of animal species by infants. *Journal of Experimental Child Psychology*, 63:189–211.

Redington, M. and Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Science*, 1(7):273–281.

Reichart, R. and Rappoport, A. (2009). The nvi clustering evaluation measure. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 165–173, Stroudsburg, PA, USA. Association for Computational Linguistics.

Riordan, B. and Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and verbal Behavior*, 14:665–681.

Rips, L. J., Smith, E. E., and Medin, D. L. (2012). Concepts and Categories: Memory, Meaning and Metaphysics. In Holyoak, K. J. and Morrison, R. G., editors, *The Oxford Handbook of Thinking and Reasoning*, pages 177–209. Oxford University Press.

Ritter, A., Mausam, and Etzioni, O. (2010). A Latent Dirichlet Allocation Method for Selectional Preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Uppsala, Sweden.

Rogers, T. T. and McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge: The MIT Press.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, pages 328–350.

Rosch, E. (1977). Studies in cross-cultural psychology. volume 1, pages 1–49. London: Academic Press.

Rosch, E. (1978). Principles of categorization. *Cognition and Categorization*, pages 27–48.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Braem, P. B. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420.

Ross, B. H. (1997). The Use of Categories Affects Classification. *Journal of Memory and Language*, 37(2):240–267.

Ross, B. H. (2000). The effects of category use on learned categories. *Memory & Cognition*, 28(1):51–63.

Roy, B. C., Frank, M. C., and Roy, D. (2012). Relating Activity Contexts to Early Word Learning in Dense Longitudinal Data. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guinness, J., Levit, M., and Gorniak, P. (2006). The human speechome project. In Vogt, P., Sugita, Y., Tuci, E., and Nehaniv, C., editors, *Symbol Grounding and Beyond: Third International Workshop on the Emergence and Evolution of Linguistic Communication*, pages 192–196. Springer Berlin Heidelberg.

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Sagi, E., Kaufmann, S., and Clark, B. (2009). Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece.

Salaberri, H., Salaberri, I., Arregi, O., and Zapirain, B. n. (2015). IXAGroupEHUDiac: A Multiple Approach System towards the Diachronic Evaluation of Texts. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 840–845, Denver, CO, USA.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 726–731.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological review*, 117(4):1144–1167.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Schyns, P. G., Goldstone, R. L., and Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, 21:1–17.

Schyns, P. G. and Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:681–696.

Shafto, P., Kemp, C., Mansinghka, V., and Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, 120(1):1 – 25.

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

Smith, E. E., Shoben, E. J., and Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81(3):214–241.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Spalding, T. L. and Ross, B. H. (2000). Concept learning and feature interpretation. *Memory & Cognition*, 28:439–451.

Starkey, D. (1981). The origins of concept formation: Object sorting and object preference in early infancy. *Child Development*, pages 489–497.

Stevenson, A., editor (2010). *The Oxford English Dictionary*. Oxford University Press, third edition.

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, 133(3):234–243.

Storms, G., Boeck, P. D., and Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, 42:51–73.

Szymanski, T. and Lynch, G. (2015). UCD: Diachronic Text Classification with Character, Word, and Syntactic N-grams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 879–883, Denver, CO, USA.

Tahmasebi, N., Risse, T., and Dietze, S. (2011). Towards automatic language evolution tracking, A study on word sense tracking. In *Proceedings of the Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn 2011)*, Bonn, Germany.

Taylor, M. and Gelman, S. A. (1988). Adjectives and nouns: Children's strategies for learning new words. *Child Development*, 59(2):411–419.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. 331(6022):1279–1285.

Theakston, A. L., Ibbotson, P., Freudenthal, D., Lieven, E., and Tomasello, M. (2015). Productivity of noun slots in verb frames. *Cognitive Science*, 39(6):1369–1395.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., and Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, 28:127–152.

Trauble, B. and Pauen, S. (2007). The role of functional information for infant categorization. *Cognition*, 105(2):362–379.

Traugott, E. and Dasher, R. (2001). *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84:327–352.

Utt, J., Springorum, S., Köper, M., and im Walde, S. S. (2014). Fuzzy V-Measure – An Evaluation Method for Cluster Analyses of Ambiguous Data. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 581–587.

Veneziano, E. (2001). Displacement and informativeness in child-directed talk. *First Language*, (21):323–356.

Vinson, D. and Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.

Vivalt, E. (2014). Modelling gaussian fields and geostatistical data using gaussian markov random fields. Technical Report. `http://evavivalt.com/wp-content/uploads/2014/11/thesis.pdf`.

Vlachos, A., Korhonen, A., and Ghahramani, Z. (2009). Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 74–82, Stroudsburg, PA, USA. Association for Computational Linguistics.

Voorspoels, W., Vanpaemel, W., and Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review*, 15(3):630–637.

Vul, E., Goodman, N., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637.

Vul, E. and Pashler, H. (2008). Measuring the crowd within: probabilistic representations within individuals. *Psychological Science*, 19(7):645–647.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). Rethinking lda: Why priors matter. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. Curran Associates, Inc.

Warrington, E. K. and Shallice, T. (1984). Category specific semantic impairments. *Brain*, 107:829–854.

Waxman, S. R. (1990). Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development*, 5(2):123 – 150.

Waxman, S. R. and Markov, D. B. (1998). Object properties and object kind: twenty-one-month-old infants' extensions of novel adjectives. *Child Development*, 69:1313–1329.

Waxman, S. R. and Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3):257–302.

Wiemer-Hastings, K. and Graesser, A. C. (2000). Contextually Representing Abstract Concepts with Abstract Structures. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pages 983–989.

Wijaya, D. T. and Yeniterzi, R. (2011). Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, pages 35–40, Glasgow, Scotland, UK.

Wu, L. and Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, 132(2):173 – 189.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, 85(3):223 – 250.

Xu, F. and Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, 114(2):245–272.

Yao, X. and Durme, B. V. (2011). Nonparametric Bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 10–14.

Younger, B. A. and Fearing, D. D. (2000). A global-to-basic trend in early categorization: Evidence from a dual-category habituation task. *Infancy*, 1(1):47–58.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, 17(3):381–397.

Yu, C. and Ballard, D. (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception*, 1:57–80.

Yu, C. and Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomput.*, 70(13-15):2149–2165.

Yu, C., Ballard, D. H., and Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, 29(6):961–1005.

Yu, C., Smith, L., Shen, H., Pereira, A., and Smith, T. (2009). Active information selection: Visual attention through the hands. *IEEE Transactions on Autonomous Mental Development*, 1(2):141 – 151.

Yu, C. and Smith, L. B. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science*, 18(5):414–420.

Yu, C. and Smith, L. B. (2010). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, 14(2):165–180.

Yu, C. and Smith, L. B. (2016). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*. early view.

Zampieri, M., Ciobanu, A. M., Niculae, V., and Dinu, L. P. (2015). AMBRA: A Ranking Approach to Temporal Text Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 851–855, Denver, CO, USA.

Zaritskii, V., Svetnik, V., and Shimelevich, L. (1976). Monte-carlo technique in problems of optimal information processing. *Automation and Remote Control*, 36(12):2015–2022.

Zeigenfuse, M. D. and Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica*, 133(3):283–295.