Incremental Bayesian Category Learning from Natural Language

Incremental Bayesian Category Learning from Natural Language

Lea Frermann and Mirella Lapata
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
l.frermann@ed.ac.uk, mlap@inf.ed.ac.uk

Author Note

We would like to thank Charles Sutton for his valuable feedback. We acknowledge the support of EPSRC through project grant EP/I037415/1.

Incremental Bayesian Category Learning from Natural Language

Models of category learning have been extensively studied in cognitive science and primarily tested on perceptual abstractions or artificial stimuli. In this paper we focus on categories acquired from natural language stimuli, that is words (e.g., *chair* is a member of the FURNITURE category). We present a Bayesian model which, unlike previous work, learns both categories and their features in a single process. 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 categories *incrementally* using particle filters, a sequential Monte Carlo method commonly used for approximate probabilistic inference which sequentially integrates newly observed data and can be viewed as a plausible mechanism for human learning. 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.[1]

Introduction

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 perspective 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

[1] An earlier version of this work was published in Frermann and Lapata (2014).

(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 exemplars), 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 article, 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`}[2]).

[2]Throughout this paper we will use small cap fonts to denote CATEGORIES, italics to denote their *members*,

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.[3] 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, by identifying ideal or optimal solutions to them and then using algorithms for approximating these solutions. 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. 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

and typewriter fonts for their `features`.

[3]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.

inference, essentially reducing probabilistic computations to generating samples from a probability distribution.

Historically, the stimuli involved in categorization studies (either lab experiments or simulations) tend to be concrete objects with an unbounded number of features (e.g., physical objects; Bornstein and Mash 2010) or highly abstract ones, with a small number of manually specified features (e.g, binary strings, colored shapes; Medin and Schaffer 1978; Kruschke 1993). 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 categorization 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 article, 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 simulations 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 exemplars created by humans.

Related Work

Theories of Categorization. Numerous theories as to how humans categorize objects have been proposed and extensively tested. It is beyond the scope of this article to provide a detailed overview, 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 exemplars encountered 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 simulations 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 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 much 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 exemplars (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 exemplars 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). 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.

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 exemplars into categories based on their features. In the remainder of this article, we use the term *exemplars* to refer to the concepts being categorized and the term *stimuli* to denote observations of exemplars and their features. A common assumption is that exemplars 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 exemplars.

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:

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

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 (1) is the prior probability of selecting category j based on the category assignments of the previously assigned exemplars. 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 (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 exemplars and features. For example, Sanborn et al. (2010) (following Medin and Schaffer 1978) use a small number of artificial exemplars, each with four binary features (e.g., 1111, 0101, 1010). In another experiment, they use 12 exemplars with continuous features, varying in brightness and saturation. Other work focusing on natural language categorization has assumed that concepts (i.e., abstract cognitive representations of exemplars) can be represented as sets of features obtained from norming studies. Table 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 1 (bottom) shows examples of the linguistic features we consider for different concepts.

The BayesCat Model

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 a cognitively plausible inference algorithm for our model.

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

$$[c_{-n} \ldots c_{-1} \quad t \quad c_1 \ldots c_n]. \tag{2}$$

Each induced category will be characterized by a set of exemplars 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 exemplars (i.e., target words) ϕ_k and (2) an independently parametrized distribution over context words ψ_k. The separation of exemplars 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 exemplars *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.

A graphical overview of the model in form of a plate diagram is presented in Figure 1(a). Observed variables (target exemplars 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 exemplars ϕ for each category k).

The generative process of the BayesCat model is displayed in Figure 1(b) and proceeds as follows.[4] 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 exemplar 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 exemplar 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 $w_c^{d,i} \sim Mult(\psi_{z^d})$.

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

The full joint distribution over data and model parameters as defined by our model (see the independence assumptions in the plate diagram in Figure 1(a)) can be factorized as:

$$P(\mathbf{y}, \mathbf{z}, \theta, \phi, \psi; \alpha, \beta, \gamma) = P(\theta|\alpha) \times \prod_{k=1}^{K} P(\phi_k|\beta)P(\psi_k|\gamma) \times \prod_{d=1}^{D} P(z^d|\theta)P(w_t^d|\phi_{z^d}) \prod_{i=1}^{I} P(w_c^{d,i}|\psi_{z^d}), \tag{3}$$

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. In particular, we can analytically integrate over all possible values of the model's parameter distributions θ, ϕ and ψ, without having to compute them explicitly. As we explain below below this model formulation allows for efficient learning.

Incremental Category Learning

In the previous section we 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 introduce a cognitively plausible learning mechanism.

A prevalent characteristic of human learning is its incrementality. 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). Equation (3) defines a probability distribution over all possible partitionings of the exemplars into categories. Exact computation of this density is both computationally intractable and 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 exemplars through a set of samples and do so in an *incremental* fashion. Each sample will correspond to one possible categorization of the observed exemplars, 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 exemplars. We achieve this by committing to past categorization decisions made by the learning algorithm, and thus integrate a new exemplar *given* the category assignments of all previously encountered exemplars (however, we will relax the strict incrementality assumption in the following section).

We develop a sequential Monte Carlo-based approximate inference algorithm for our model. Monte Carlo (MC) methods approximate complex densities through a set of random samples from those densities.[5] While most such methods operate in batch mode, requiring the availability of all input data before learning, some *sequential Monte Carlo* methods have been developed, where samples from the target posterior distribution are updated incrementally as more information becomes available over time. In the following section we illustrate our learning algorithm schematically using the example in Figure 2a; we refer the interested reader to Appendix B for a more technical description.

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

[5] With the number of random samples approaching infinity, the approximation is guaranteed to converge towards the target distribution.

process by treating the observation of each stimulus as one time point. In the example in Figure 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.

During learning, we incrementally approximate the target density, i.e., the probability distribution over all possible categorizations of all exemplars $p_T(\boldsymbol{z}_{1:T}|\boldsymbol{y}_{1:T})$ through a cascade of local importance distributions $p_t(\boldsymbol{z}_{1:t}|\boldsymbol{y}_{1:t})$. At each time t, p_t is the distribution over clusterings $\boldsymbol{z}_{1:t}$ of observed exemplars $\boldsymbol{y}_{1:t}$, represented through the current set of particles. Figure 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 SIS framework, we present a proposal distribution $q(\cdot)$ where we assume that once an exemplar has been assigned a category, this category is *fixed*:

$$q_t(\boldsymbol{z}_{1:t}|\boldsymbol{y}_{1:t}) = q_{t-1}(\boldsymbol{z}_{1:t-1}|\boldsymbol{y}_{1:t-1})q_t(z_t|z_{t-1}, y_t). \tag{4}$$

Here, 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 exemplar 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 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. We use the posterior distribution:

$$q_t(z_t|z_{t-1}, y_t) = p(z_t|\boldsymbol{z}_{1:t-1})p(y_t|z_t), \tag{5}$$

taking into account prior information about category probability and the features of the exemplar. 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 (please refer to Appendix B for more information).

By repeatedly sampling from local approximations to the target density, inaccuracies will inevitably accumulate. 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. We incorporate a threshold-based resampling scheme, measuring weight variance as *effective sample size* (ESS):

$$ESS(\boldsymbol{w}^t) = \left(\frac{1}{\sum_n (w_n^t)^2}\right) \tag{6}$$

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. Technically, resampling consists of drawing N times with replacement from a multinomial distribution over particles parametrized by the current set of particle weights. Figure 2a shows one resampling step following the particle updates. In the resampling step, 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).

Relaxing Strict Incrementality

The learning algorithm presented above approximates the target distribution over categorizations of observed exemplars 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 re-sampling step for each particle, we individually reconsider the category assignment for a fixed number of previously encountered exemplars. 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 2a illustrates rejuvenation for the bottom set of particles. Each particle revisits one previous categorization decision (e.g., the blue particle, places exemplar 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.

Modeling Experiments

In the following sections we present a series of simulations assessing the performance of the model presented above on a category acquisition task. Our simulations are designed to examine whether the model produces meaningful categories but also to investigate the

learning process itself and its characteristics. In the first simulation 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. Simulations 2 and 3 explore category acquisition in children using a corpus of child-directed speech, whereas Simulation 4 presents a typicality rating simulation. All our simulations evaluate the categories produced by the models against gold standard categories created by humans.

Simulation 1: Large-scale Category Acquisition

Our first goal was to examine whether any meaningful categories emerge when our incremental model is trained on a large corpus. We compare the BayesCat model 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.[6] 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 datasets contain a classification of (concrete) nouns into (possibly multiple) semantic categories produced by human participants. Examples from the dataset are provided in Table 2. The former dataset 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

[6]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/

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 dataset, 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-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) dataset with the concepts used in Vinson and Vigliocco (2008). The latter dataset 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).[7] The size and nature of this evaluation dataset is in sharp contrast to those used in previous categorization studies which consist of a small number of artificial

[7]The dataset is available from `www.frermann.de/data`.

concepts.

The input to all models comprises the same set of linguistic stimuli, each of which consists of one target word t, surrounded by a symmetric window of n context words $[c_{-n} \ldots c_{-1} \;\; t \;\; c_1 \ldots c_n]$. The target words are defined by the set of concepts included in our gold standard. Some corpus statistics are given in Table 3 (column BNC). The corpus was lemmatized and stopwords were removed. Infrequent context words (occurring less than 800 times) were also eliminated. 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. Exemplars 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 — exemplars are 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. Exemplars 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 exemplar. 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 simulations.

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. Specifically, we adopted a Gibbs sampler as a batch learning strategy for our BayesCat model. Gibbs sampling (Geman and Geman, 1984) is a Markov chain Monte Carlo technique for approximating complex joint probability distributions. The model parameters, are initialized at random, and the sampler performs multiple sweeps

over the set of stimuli, until convergence. The joint probability density is approximated by repeated re-sampling from the conditional density of individual latent labels given the current assignment of all other latent variables. 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 simulations 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:

$$cat(w) = \max_z \; P(w|z) \cdot P(z|w) \tag{7}$$

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:

$$\text{pu} = \frac{1}{N} \sum_i \max_j |C_i \cap G_j| \tag{8}$$

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:

$$\text{co} = \frac{1}{N} \sum_j \max_i |C_i \cap G_j| \tag{9}$$

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, homogeneity is weighted more strongly:

$$\text{F}_\beta = \frac{(1+\beta) \cdot pu \cdot co}{(\beta \cdot pu) + co} \tag{10}$$

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.[8] 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):

$$\text{VH} = 1 - \frac{H(G|C)}{H(G)} \tag{11}$$

$$\text{VC} = 1 - \frac{H(C|G)}{H(C)} \tag{12}$$

$$\text{VM} = 1 - \frac{(1+\beta) \cdot VH \cdot VC}{(\beta \cdot VH) + VC} \tag{13}$$

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

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

various entropy values involve the estimation of the joint probability of induced class C and gold standard class G:

$$\hat{p}(C, G) = \frac{\mu(C \cap G)}{N} \tag{14}$$

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 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 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.

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, 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 exemplars and more accurate clusterings can be (on average) achieved more easily.

Table 5 shows example categories learnt by the incremental BayesCat model. Each induced category is characterized by a set of exemplars (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 3 and 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 build on the basis of co-occurrence counts of target words and context words. With an increasing number of observations, however, these counts become less distinctive between target concepts. Inspection of the output of the CW algorithm, reveals that it induces one very big category, 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.

Discussion. In this simulation, 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 substantially 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 3a). For the incremental BayesCat model, however, the purity of categories improves constantly as well as well their completeness (see Figures 3a and 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).

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.

Simulation 2: Child Category Acquisition

The primary goal of the preceding experiment was to explore how effectively our model captures large-scale category information. Of greater 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 simulation 1. To answer this question, we applied our incremental model to a corpus of child-directed speech and evaluated the resulting categories against the gold standard clusters used previously.

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 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 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 exemplars occurring 50 times or more within the corpus. Analogously, we filtered low-frequency context words with the same threshold. 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 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), and incremental Chinese Whispers (CW). 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

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

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

evaluation metrics. Examples of the semantic categories induced by BayesCat are shown in Table 7.

Figures 5 and 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 expense of purity. This effectively means that CW tends to learn one very big cluster comprising of the majority of exemplars. PF, on the other hand, shows clear learning curves across metrics, with increasingly clean (Figures 5(a) and 6(a)) and complete clusters (Figures 5(b) and 6(b)).

Discussion. In this simulation we showed that the BayesCat model can learn meaningful categories from a corpus of child-directed speech. Compared to the previous simulation, 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.

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 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 exemplars from this category in the presence of young children. In contrast, the WEAPONS category is acquired from the BNC (see Table 5), which, again, emphasizes the ability of our model to adapt to and learn from empirical data.

Table 7 provides qualitative examples of the categories and features learnt by BayesCat. As can be seen, categories are coherent and easily interpretable, with relevant features. 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.

Simulation 3: Memory Constraints

In this simulation we delve deeper into our incremental inference algorithm and its appropriateness for human, 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 simulations, 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 simulations. 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. 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 8a,b 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 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 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 simulation was too small.

Figure 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 9b).

Figure 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 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, hereby confirming our intuition that a learner's knowledge state should become increasingly confident over time, and reconsiderations of past decisions should decrease in frequency.

Discussion. In this simulation we compared the effect of memory resources on the learning behavior of the incremental BayesCat model by examining 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. Figure 10 demonstrates that in the initial learning phase resampling is very frequent, but the frequency decreases over time.

Simulation 4: 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 exemplars of a category and must predict the degree to which the exemplars 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 an exemplar is derived by summing the similarity of that exemplar to all exemplars in the category. More formally, the typicality of exemplar w for category G is given by:

$$T_G(w) = \sum_{v \in G} \eta_{w,v} \tag{15}$$

where $\eta_{w,v}$ is the similarity of exemplar w to exemplar v, with v also belonging to category G. The similarity function $\eta_{w,v}$ can vary depending on how exemplars 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):

$$\begin{aligned} \eta_{w,v} = P(v|w) &= \sum_k P(v|k)P(k|w) \\ &= \sum_k P(v|k)\frac{P(w|k)P(k)}{P(w)} \end{aligned} \tag{16}$$

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

In this simulations, 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 simulations, we transformed raw co-occurrence counts into PMI values. A typicality value for each member of a category was computed

using (15) and summing the cosine similarity of the exemplar vector $\overrightarrow{w}$ to the all other vectors representing its co-members $\overrightarrow{v}$:

$$\eta_{w,v} = cos\Big(\overrightarrow{w}, \overrightarrow{v}\Big) = \frac{\overrightarrow{w} \cdot \overrightarrow{v}}{|\overrightarrow{w}||\overrightarrow{v}|} \tag{17}$$

Our simulations used the dataset produced by Fountain and Lapata (2010) who elicited typicality ratings[11] (and category labels) for all exemplars contained in the feature norms 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 dataset. 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 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 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 8 shows some qualitative examples of concepts which BayesCat rated as most typical/atypical for a particular category.

Discussion. In this set of simulations we compared two models in their ability to rank exemplars with respect to typicality, against a human created gold standard. We

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

showed that our model successfully captured the typicality of exemplars within a given category. As can be observed in Table 8, many of the typicality ratings produced by BayesCat correspond to human intuitions. We should also point out that this is a large scale study over hundreds of exemplars. 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.

General Discussion

In this paper we have presented 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 simulations 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 simulation showed that our model performs well on a typicality rating task when compared against a non-incremental semantic space.

In our simulations, 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. The Gibbs sampler and the incremental learner acquire categories from identical corpora. The Gibbs sampler,

however, can make optimal use of the information encoded in the corpus, whereas our incremental learner has limited access to the training data. In view of this limitation, the incremental particle filters perform competitively throughout our simulations.

BayesCat has a cognitively plausible learning mechanism and induces meaningful categories. However, it learns a flat set of features, even though there is evidence suggesting that humans organize their category knowledge hierarchically (Palmeri, 1999; Verheyen et al., 2008). Furthermore, our model acquires features individually for each category. For example, it does not learn that ANIMALS can be described in terms of their `behavior` and `diet`, whereas FURNITURE or TOOLS cannot. On a related note, the model learns unstructured bags-of-features even though it has been shown that humans learn features that are shared across categories (Ahn, 1998; Spalding and Ross, 2000). In the future, we would like to devise more sophisticated models of categorization which jointly learn categories and feature types (e.g., `behavior`). 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 simulations, 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.

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 simulations 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 simulations 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 article has been primarily on the learning mechanisms of categorization, our simulations 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.

References

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.

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.

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.

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, New York City.

Blei, D. M., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *JMLR*, 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, Canberra, Australia.

Börschinger, B. and Johnson, M. (2012). Using rejuvenation to improve particle filtering for bayesian word segmentation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–89, Jeju Island, Korea.

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, Athens, Greece.

Brown, S. D. and Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58:49–67.

Canini, K. (2011). *Nonparametric Hierarchical Bayesian Models of Categorization*. PhD thesis, EECS Department, University of California, Berkeley.

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.

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, Austin, TX: Cognitive Science Society.

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, Austin, TX, USA.

Cree, G., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, 23(3):371–414.

Daw, N. D. and Courville, A. (2007). The pigeon as particle filter. In *Advances in Neural Information Processing Systems*, volume 20, pages 369–376, Cambridge, MA. MIT Press.

Diaz, M. and Ross, B. H. (2006). Sorting out categories: Incremental learning of category structure. *Psychonomic Bulletin and Review*, 13(2):251–256.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

Evans, J. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Essays in cognitive psychology. Psychology Press.

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, Nice, France.

Fountain, T. and Lapata, M. (2010). Meaning representation in natural language categorization. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 1916–1921, Portland, Oregon. Cognitive Science Society.

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, Boston, Massachusetts.

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.

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*, pages 249–258, Gothenburg, Sweden.

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.

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.

Goldstone, R. L., Lippa, Y., and Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, 78:27–43.

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.

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, Austin, TX, USA.

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., Tenenbaum, J. B., and Steyvers, M. (2007b). Topics in semantic representation. *Psychological Review*, 114:2007.

Hammersley, J. M. and Morton, K. W. (1954). Poor Man's Monte Carlo. *Journal of the Royal Statistical Society. Series B*, 16(1):23–38.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(23):146–162.

Heit, E. and Barsalou, L. W. (1996). The instantiation principle in natural categories. *Memory*, 4(4):413–451.

Hol, J. D., Schön, T. B., and Gustafsson, F. (2006). On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop*.

Jern, A. and Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, 66:85–125.

Jones, S. S., Smith, L. B., and Landau, B. (1991). Object properties and konwledge in early lexical learning. *Child Development*, 62(3):499–516.

Kemp, C., Shafto, P., and Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, 64:35–75.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, 5:3–36.

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, Edinburgh, Scotland, UK.

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.

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.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third Edition.* Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

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.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and non-living things. *Behavioral Research Methods Instruments & Computers*, 37(4):547–559.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

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.

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. (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.

Palmeri, T. J. (1999). Learning categories at different hierarchical levels: a comparison of category learning models. *Psychonomic Bulletin & Review*, 6:495–503.

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.

Riordan, B. and Jones, M. (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.

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, chapter Human Categorization, pages 1–49. London: Academic Press.

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, Prague, Czech Republic.

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

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., Navarro, D. J., and Griffiths, T. L. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167.

Schyns, P. G. and Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:681–696.

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, Honolulu, Hawaii. 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.

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.

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 –, Reykjavik, Iceland.

Verheyen, S., Ameel, E., Rogers, T. T., and Storms, G. (2008). Learning a hierarchical organization of categories. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 751–757, Austin, TX, USA.

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.

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.

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, Portland, Oregon.

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.

Zeigenfuse, M. and Lee, M. (2010). Finding the features that represent stimuli. *Acta Psychologica*, 133(3):283–295.

		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 1

Exemplars and their features for the categories FRUIT *and* ANIMAL. *Features are shown as feature norms (top) and as context words (bottom).*

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 2

Example categories and their concepts taken from our gold standard.

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

Table 3

Number of stimuli, exemplars, and features retrieved from BNC and CHILDES.

	Development Set					
	pu	co	$F_{0.5}$	VH	VC	VM
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
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

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

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 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).

	pu	co	$F_{0.5}$	VH	VC	VM
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 6

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

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, jam

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, van, break

Table 7

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

(a) CHILDES

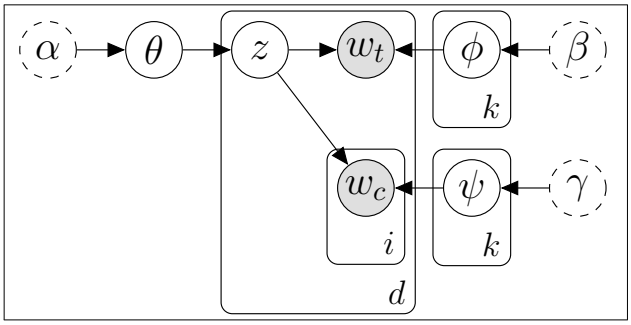
category	most typical concepts	least typical concepts
FOOD	*cake, bread*, strawberry, cheese*	*owl*, lobster, snail*, deer**
ANIMAL	*elephant, horse, cow*, duck*	*bat, pickle, chipmunk, tuna**
CLOTHING	*shirt*, shoe, sock, dress**	*necklace*, cap, cape, hose**
VEHICLE	*car, train*, truck*, bus**	*ship,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

(b) BNC

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 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.*



(a)

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)

Figure 1. (a) Plate diagram representation of the BayesCat model. (b) The generative process of the BayesCat model.

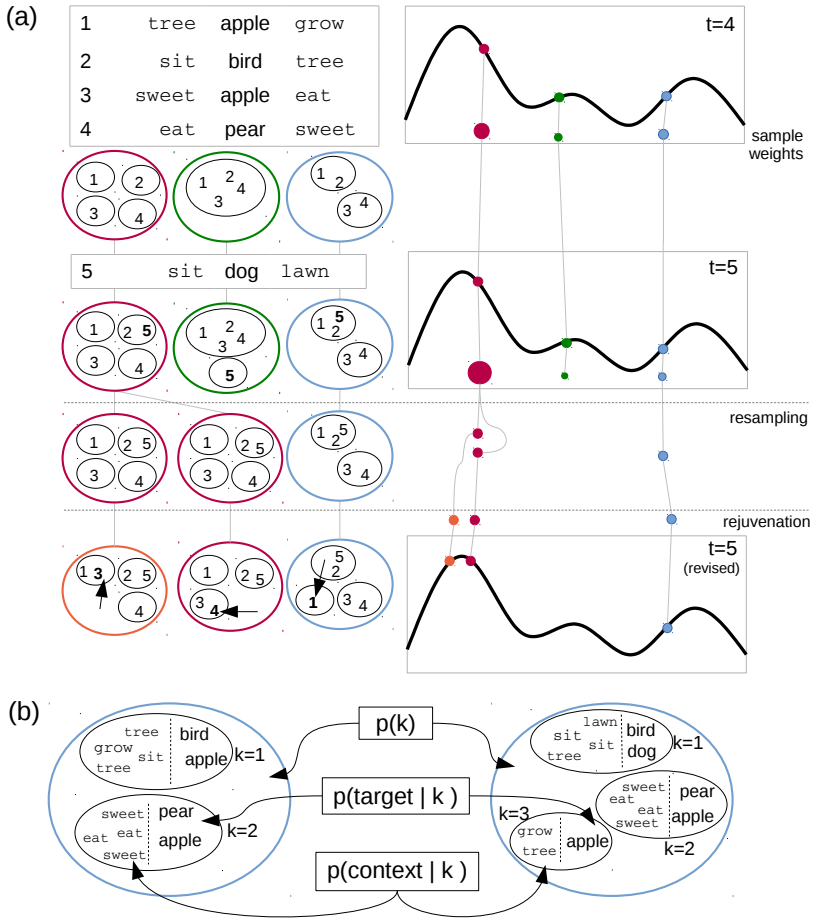
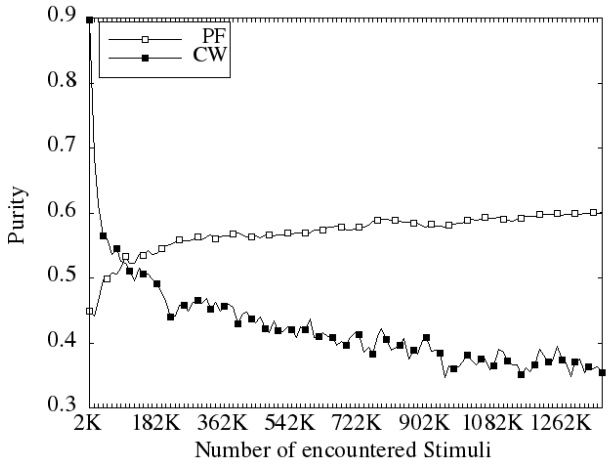
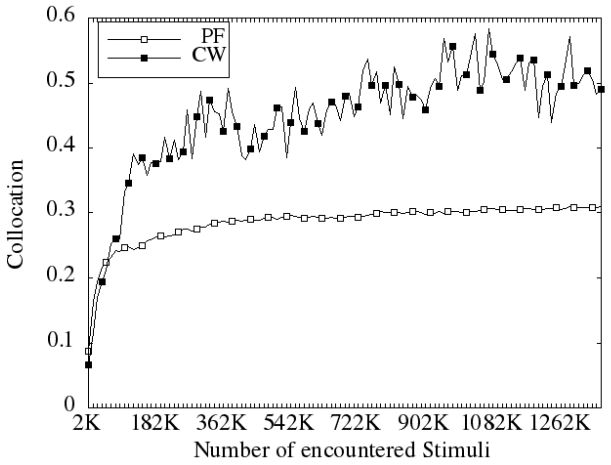


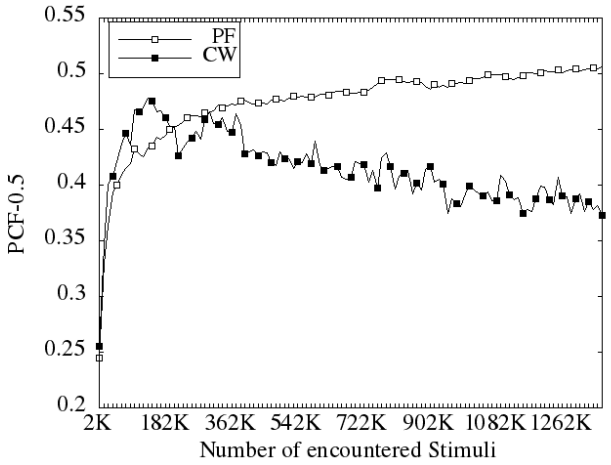
Figure 2. (a) Visualization of the particle filtering procedure in the BayesCat model using an example of a 3-particle filter. Each particle corresponds to a clustering of the observed stimuli up to time t (left), and the collection of weighted particles serves as the current approximation of the posterior distribution over clusterings (right). The 5 exemplars observed by the filter are shown in the tables. We show one update step for all particles with exemplar 5, and one subsequent re-sampling 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 the APPLE exemplar 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.



(a)



(b)



(c)

Figure 3. Learning curves for PF and CW on the BNC using purity, collocation, and $F_{o.5}$.

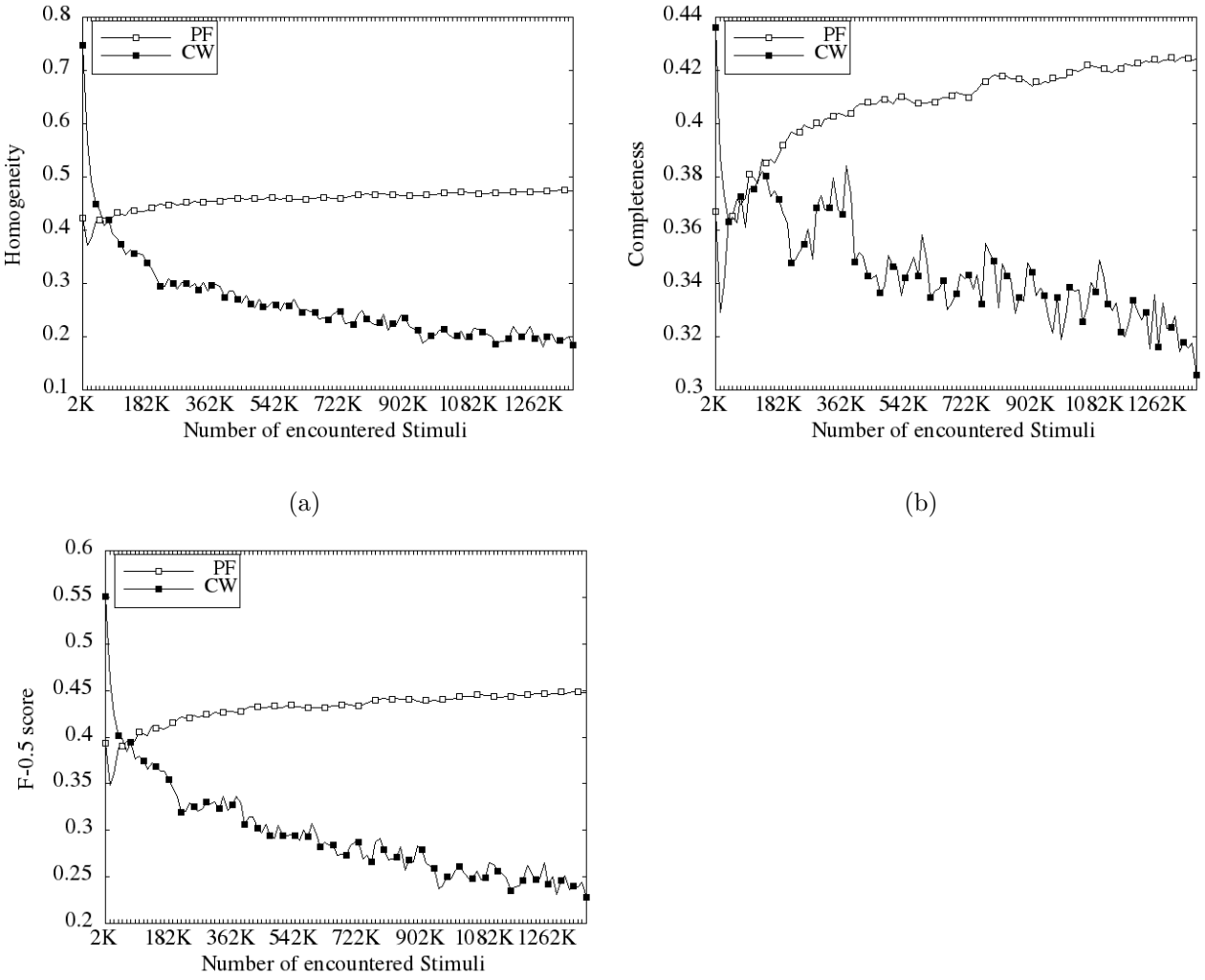


Figure 4. Learning curves for PF and CW on the BNC using (fuzzy) homogeneity, completeness, and V-measure.

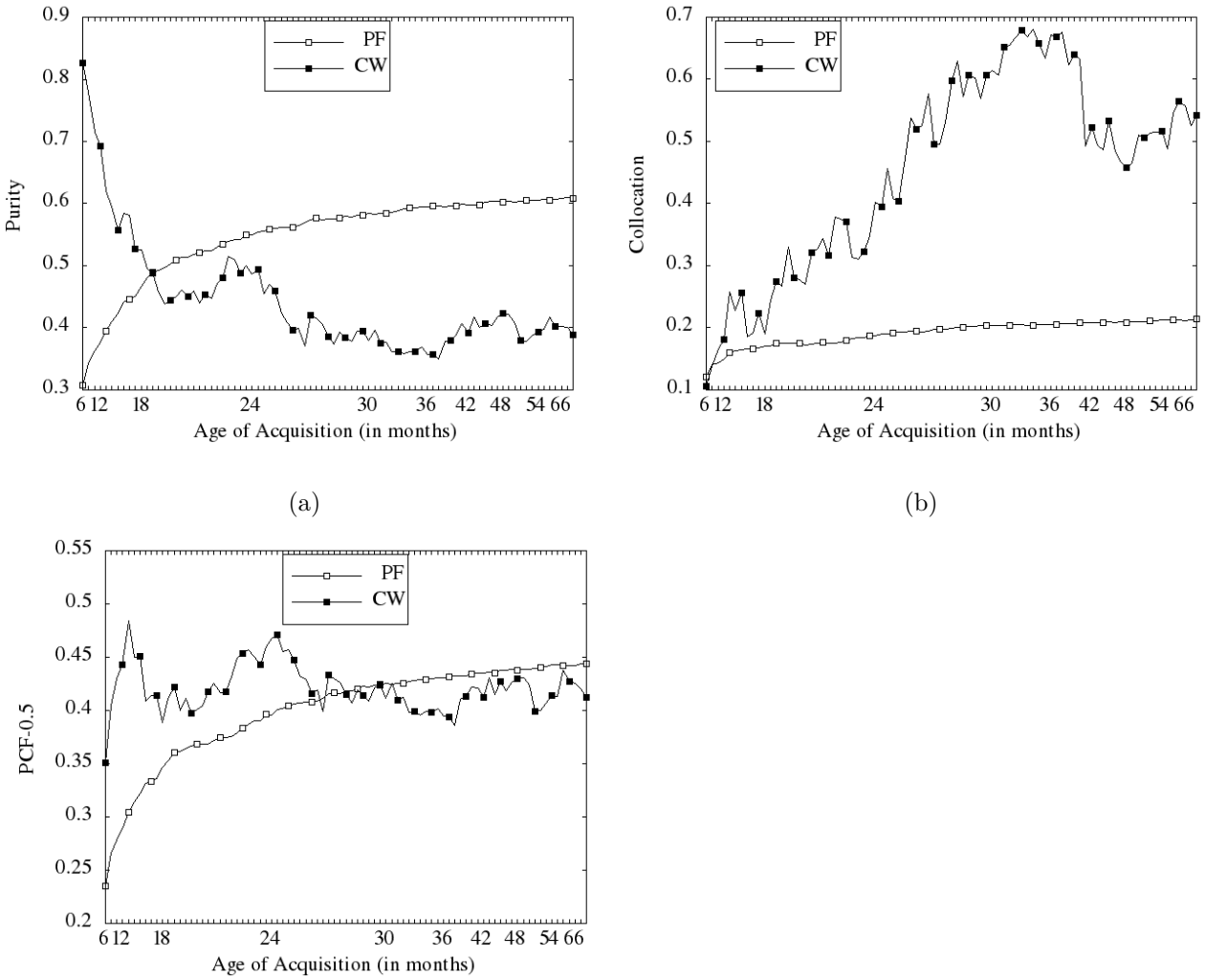


Figure 5. Learning curves for PF and CW on the CHILDES corpus using purity, collocation, and $F_{0.5}$.

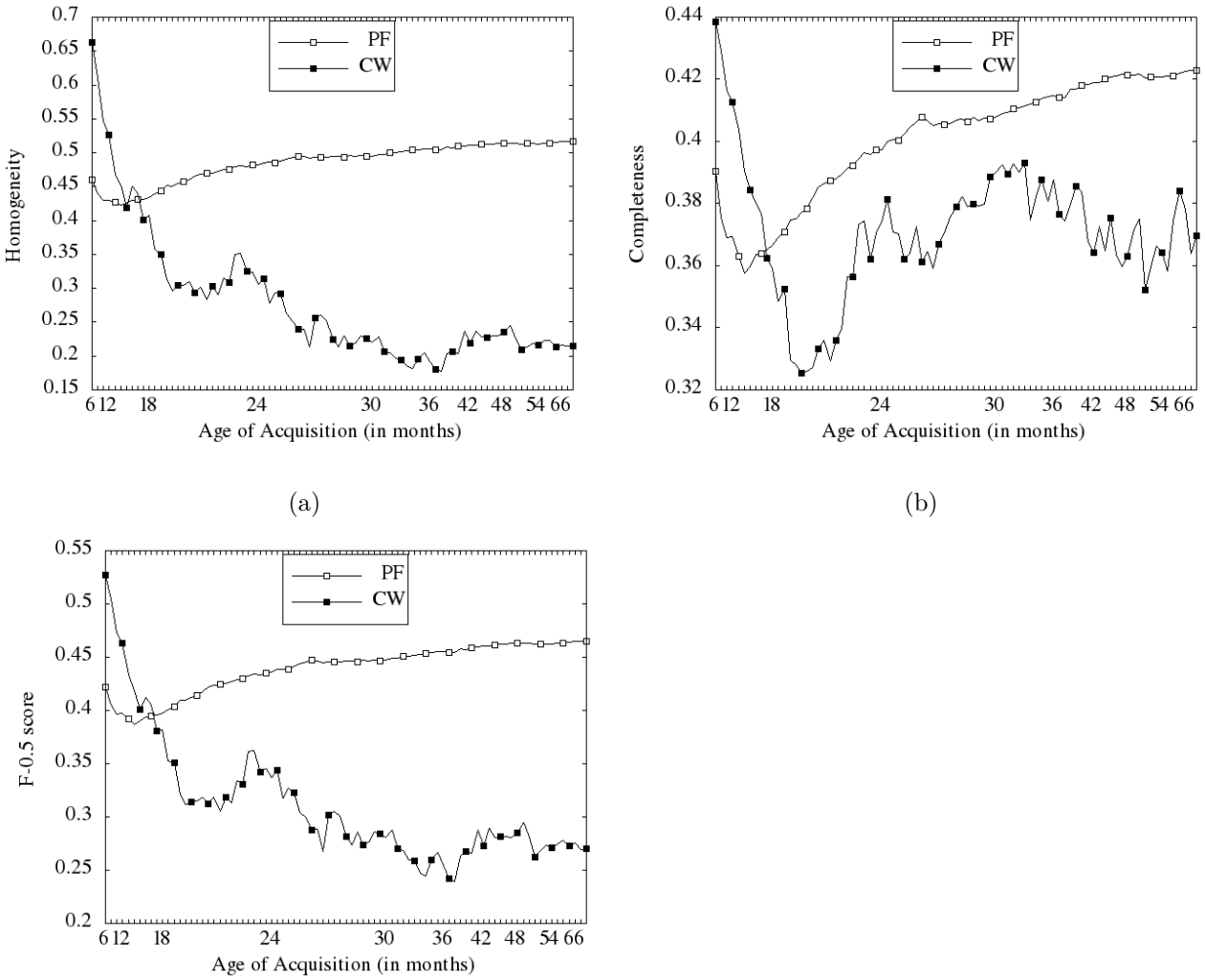


Figure 6. Learning curves for PF and CW on the CHILDES corpus using (fuzzy) homogeneity, completeness, and V-measure.

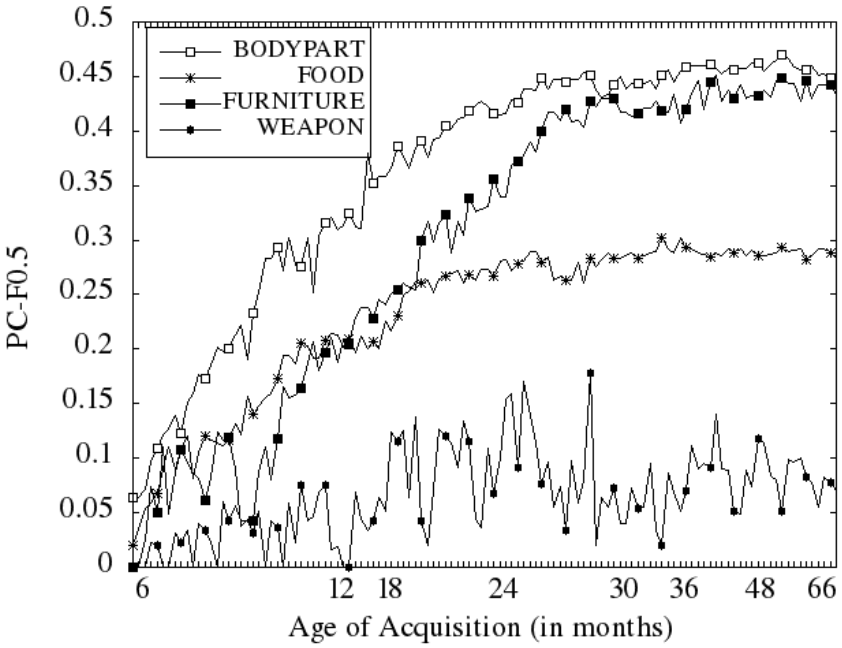
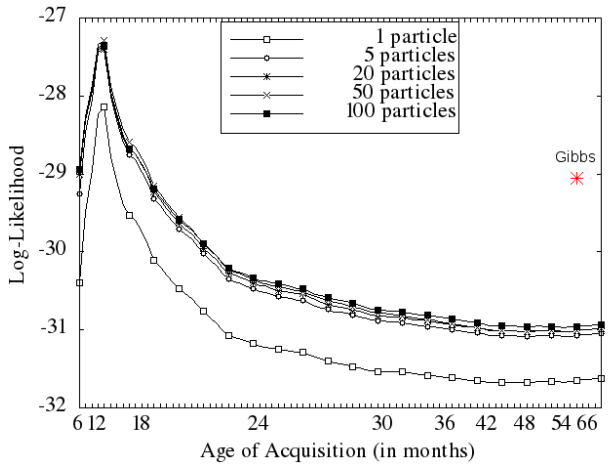
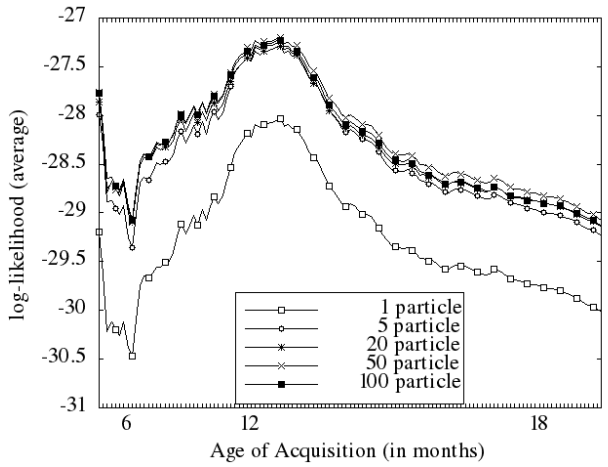


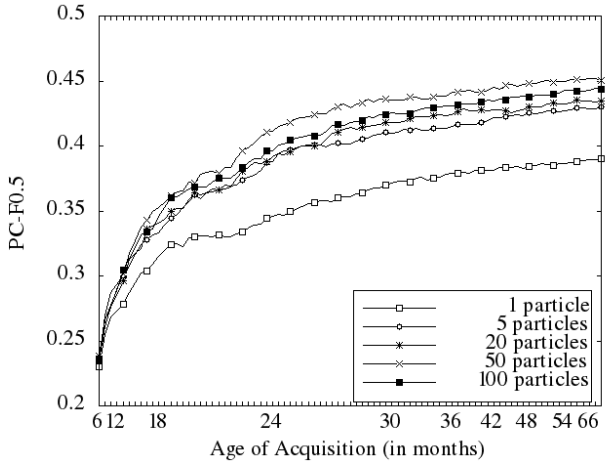
Figure 7. Emergence of selected categories over time for the incremental BayesCat model on the CHILDES corpus.



(a)

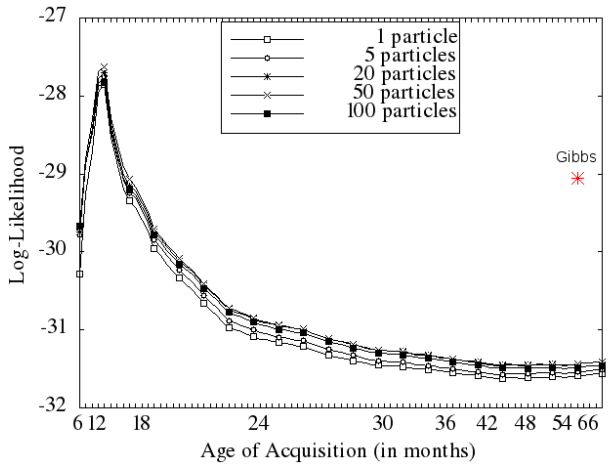


(b)

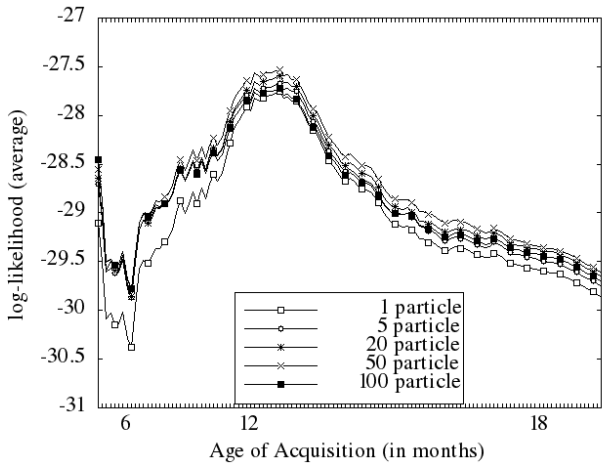


(c)

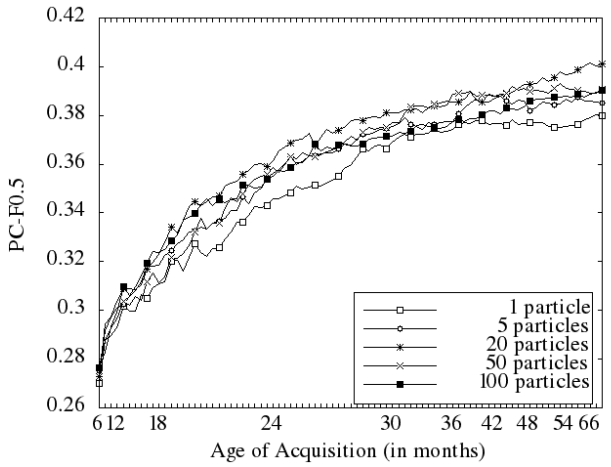
Figure 8. Learning curve for the BayesCat model on CHILDES with varying number of particles. Model log-likelihood curve (a), model log-likelihood curve for the early learning phase (b), and $F_{0.5}$ learning curve (c).



(a)



(b)



(c)

Figure 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).

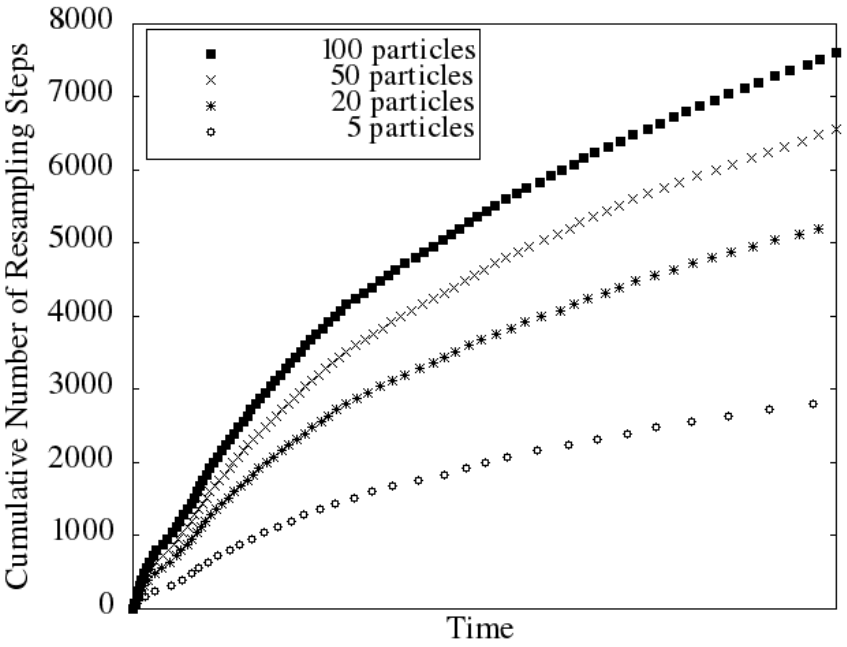


Figure 10. Resampling behavior of the BayesCat model learnt with a varying number of particles. Points correspond to executed resampling steps at time x.

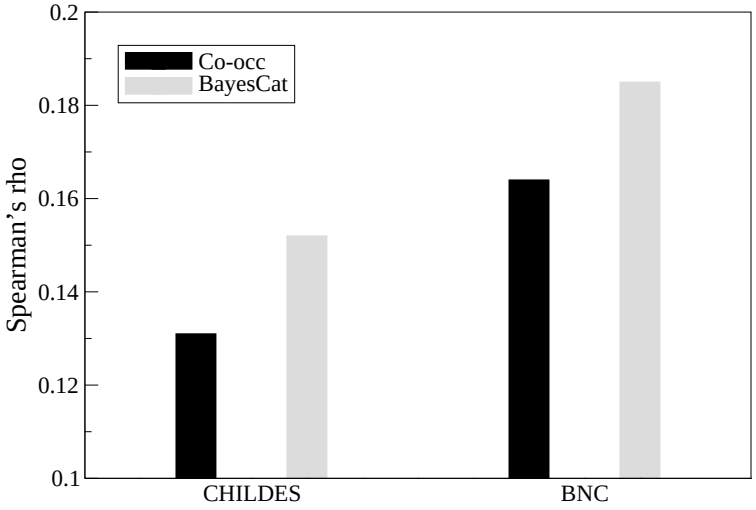


Figure 11. Rank correlations (Spearman's rho) between the gold typicality ranking and the model produced rankings over the set of all gold standard categories.

Appendix A

Details of the BayesCat Model

The full joint distribution over data and model parameters as defined by our model can be factorized as:

$$P(\mathbf{y}, \mathbf{z}, \theta, \phi, \psi; \alpha, \beta, \gamma) = P(\theta|\alpha) \times \prod_{k=1}^{K} P(\phi_k|\beta)P(\psi_k|\gamma) \times \prod_{d=1}^{D} P(z^d|\theta)P(w_t^d|\phi_{z^d}) \prod_{i=1}^{I} P(w_c^{d,i}|\psi_{z^d}), \tag{18}$$

where $\mathbf{y}$ refers to all observed data, and $\mathbf{z}$ 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. In particular, we can analytically integrate over all possible values of the model's parameter distributions θ, ϕ and ψ, without having to compute them explicitly. This is possible because their prior distributions, the Dirichlet distributions, are conjugate to the multinomial distribution, and can thus be updated in a straightforward way with new observations. Dirichlet 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 Dirichlet distribution is a commonly used prior for multinomial parameters, because of its mathematical convenience and straightforward interpretablility. It is the *conjugate prior* of the multinomial distribution, which means that the posterior distribution, resulting from their combination, has again the form of a Dirichlet distribution. Prior parameters can then be interpreted as "pseudo counts" which, during inference, are efficiently updated with counts of observations from the data. The simplified posterior distribution is:

$$P(\mathbf{y}, \mathbf{z}, \theta, \phi, \psi; \alpha, \beta, \gamma) \propto \frac{\prod_k \Gamma(\mathcal{N}_k + \alpha_k)}{\Gamma(\sum_k \mathcal{N}_k + \alpha_k)} \times \prod_{k=1}^{K} \frac{\prod_r \Gamma(\mathcal{N}_r^k + \beta_r)}{\Gamma(\sum_r \mathcal{N}_r^k + \beta_r)} \times \prod_{k=1}^{K} \frac{\prod_s \Gamma(\mathcal{N}_s^k + \gamma_s)}{\Gamma(\sum_s \mathcal{N}_s^k + \gamma_s)}, \tag{19}$$

where r ranges over target exemplars, 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 (19). Instead, the model is represented purely through occurrence counts of categories $\mathcal{N}_k$ as well as co-occurrence counts of categories with exemplars and features, $\mathcal{N}_r^k$ and $\mathcal{N}_s^k$, respectively.

Appendix B

The incremental Learning Algorithm

We first explain the sequential importance sampling procedure on which our learning algorithm is based and then derive a particle filter for the BayesCat model. Figure B1 summarizes the learning algorithm.

Importance sampling (Hammersley and Morton, 1954) is a Monte Carlo technique used to approximate a complex target distribution $p(z)$ from which samples cannot be obtained efficiently. Instead, a simpler proposal, or importance, distribution $q(z)$ is employed which is similar to the target function, but easier to sample from. The target distribution is approximated through $n = [1..N]$ samples from the importance distribution. Each sample is weighted in order to adjust for the inevitable error introduced by sampling from an approximation:

$$\begin{array}{lll} \text{Target Approximation} & \text{Sample} & \text{Weight} \\ p(z) \propto \frac{1}{N} \sum_{n=1}^{N} w^{(n)}(z^{(n)}) & z^{(n)} \sim q(\cdot) & w^{(n)} = \frac{p(z^{(n)})}{q(z^{(n)})} \end{array} \tag{20}$$

Sequential Importance sampling (SIS; Gordon et al. 1993) is an incremental version of the importance sampling algorithm. Samples from the importance distribution, as well as their weights are updated recursively with new information. A particle filter is a sequential Monte Carlo algorithm which builds on sequential importance sampling in order to incrementally approximate a target distribution (Doucet et al., 2001). In particular, a set of weighted samples, called particles, obtained through importance sampling are propagated through time $t = [1..T]$, where each sample depends on the previous samples $(1 : t-1)$:

$$\begin{array}{lll} p_T(z_{1:T}) & \propto \frac{1}{N} \sum_{n=1}^{N} w_T^{(n)}(z_T^{(n)}) & \text{Final target approximation at time T} \\ z_t^{(n)} & \sim q_t(\cdot | z_{1:t-1}^{(n)}) & \text{Sample update at time } t \\ w_t & = w_{t-1}^{(n)} \times \frac{p_t(z_t^{(n)})}{p_{t-1}(z_{t-1}^{(n)}) q_t(z_t^{(n)})} & \text{Weight update at time } t \end{array} \tag{21}$$

The set of particles at any time is a Monte Carlo approximation of the target distribution.

1: Initialize particles by randomly partitioning first d stimuli ▷ Initialization
2: Initialize weights $\mathbf{w^1} = \frac{1}{N}$
3: **for** stimulus t = [d+1...T] **do**
4: **for** particle n = [1...N] **do**
5: $z_n^t \sim q_{t-1}(\boldsymbol{z}_{1:t-1}|\boldsymbol{y}_{1:t-1})q_t(z_t|z_{t-1}, y_t)$ ▷ Particle Update
6: $S_n^t \rightarrow (S_n^{t-1}, z_n^t)$
7: $w_n^t = w_{n-1}^t * P(y^t|z^{t-1}))$ ▷ Weight Update
8: $\tilde{\mathbf{w}}^t \leftarrow normalize(\mathbf{w}^t)$
9: **if** $ESS(\tilde{\mathbf{w}}^t) \leq thresh$ **then** ▷ Resampling
10: $\mathcal{P}(i) \leftarrow \{Mult(\tilde{\boldsymbol{w}}^t)\}_{i=1}^N$
11: $\mathbf{w^t} = \frac{1}{N}$
12: **for** particle n $\in \mathcal{P}(i)$ **do** ▷ Rejuvenation
13: **for** stimulus o=[1...O] **do**
14: $d^o \sim uniform(1 \dots t)$
15: $z_n^{d^o} \sim P(z_n^{d^o}|\boldsymbol{z}_{n\setminus -d^o}^t, \boldsymbol{y}_t)$

Figure B1. The particle filtering procedure.

During learning of the BayesCat model, we incrementally approximate the target density, i.e., the probability distribution over all possible categorizations of all exemplars $p_T(\boldsymbol{z}_{1:T}|\boldsymbol{y}_{1:T})$ through a cascade of local posterior probability distributions $p_t(\boldsymbol{z}_{1:t}|\boldsymbol{y}_{1:t})$. At each time t, p_t is the distribution over clusterings $\boldsymbol{z}_{1:t}$ of observed exemplars $\boldsymbol{y}_{1:t}$, represented through the current set of particles. In order to compute the exact posterior distribution, the categorization of exemplar $\boldsymbol{y}_{1:t-1}$ would need to be re-computed for each time step considering all observed evidence. Regarding our BayesCat model, this would come with the advantage that categorizations which receive low probability in the early

training phase, but become likely in the later training phase, can be considered. However, the exact local posterior distribution is 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 exemplars. 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.

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 exemplars. In particular, we assume that once an exemplar has been assigned a category, this category is fixed:

$$\begin{aligned} q_t(\boldsymbol{z}_{1:t}|\boldsymbol{y}_{1:t}) &= q_1(z_1|y_1)\prod_{k=2}^{t} q_k(z_k|\boldsymbol{z}_{1:k-1},\boldsymbol{y}_{1:k}) \\ &= q_{t-1}(\boldsymbol{z}_{1:t-1}|\boldsymbol{y}_{1:t-1})q_t(z_t|\boldsymbol{z}_{1:t-1},\boldsymbol{y}_{1:t}) \\ &= q_{t-1}(\boldsymbol{z}_{1:t-1}|\boldsymbol{y}_{1:t-1})q_t(z_t|z_{t-1},y_t), \end{aligned} \tag{22}$$

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 the pseudocode given in Figure B1.

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 exemplar y_t from its posterior distribution over categories:

$$q_t(z_t|z_{t-1},y_t) = p(z_t|\boldsymbol{z}_{1:t-1})p(y_t|z_t), \tag{23}$$

taking into account prior information about category probability and the features of the exemplar. The posterior distribution can be shown to be a locally-optimal choice in that it minimizes the variance of the importance weights across samples.[12] The incremental importance weights then correspond to the predictive likelihood of the current stimulus y_t:

[12]The proposal distribution $q(x)$ can be selected at liberty, as long as its support includes the support of the target distribution $p(x)$. A common choice of the proposal density in Bayesian modeling is the prior

$$
\begin{aligned}
w_t(\boldsymbol{z}_{1:t}|\boldsymbol{y}_{1:t}) &= w_{t-1} \times \frac{p(\boldsymbol{z}_{1:t}|\boldsymbol{y}_{1:t})}{q(\boldsymbol{z}_{1:t}|\boldsymbol{y}_{1:t})} \\
&\propto w_{t-1} \times p(y_t|z_{t-1}) \\
&= w_{t-1} \times \sum_{z_t} p(z_t|z_{t-1}) p(y_t|z_t, z_{t-1}).
\end{aligned} \tag{24}
$$

The weights are normalized to sum to one after each iteration (see lines 7–9 in Figure B1). Because of our compact model formulation, purely in terms of sufficient statistics, we are able to sample from the local posterior distributions, as well as to evaluate the predictive likelihood, and can thus use the optimal proposal function in our particle filter.

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. *Resampling* is one common approach to this problem: low-weight particles are replaced with copies of high-weight particles based on some pre-determined schedule. This way, memory resources can be allocated on high-probability particles, individual copies of which can be further propagated. We follow a threshold-based resampling scheme measured by the variance across the current particle weights. A commonly used measure for weight variance is the *effective sample size* (ESS):

$$
ESS(\boldsymbol{w}^t) = \left(\frac{1}{\sum_n (w_n^t)^2} \right) \tag{25}
$$

A resampling step is executed whenever the ESS falls below a set threshold.

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 10–12 in Figure B1). The resulting set of particles is an empirical estimate of the current approximation, in that the weights

distribution. In this case, the weight updates can be shown to be the likelihood function. Intuitively, it is clear that updating hypotheses purely on the basis of prior category assignments, while ignoring the features of the current exemplar, will result in noisy clusterings, and is cognitively implausible.

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).

Rejuvenation. Finally, we employ rejuvenation in order to relax the incrementality assumption of our learning algorithm Technically, rejuvenation involves, individually for each particle, the construction of a Markov transition kernel which is invariant with respect to the target distribution. Each particle is then independently moved according to the kernel and, by its definition, after the move the particles are still distributed according to the importance distribution. We instantiate our kernel as a Gibbs sampler which resamples one variable conditioned on the current values of all other variables. For a fixed and constant number of iterations we randomly select an exemplar uniformly from all encountered exemplars and resample its category based on all other current category assignments (see lines 1–16 in Figure B1).