Bootstrapping Language Acquisition

Omri Abend and Tom Kwiatkowski and Nathaniel J. Smith
and Sharon Goldwater and Mark Steedman

Institute of Language, Cognition and Computation
School of Informatics
University of Edinburgh [1]

Abstract

The semantic bootstrapping hypothesis proposes that children acquire their native language through exposure to sentences of the language paired with structured representations of their meaning, whose component substructures can be associated with words and syntactic structures used to express these concepts. The child's task is then to learn a language-specific grammar and lexicon based on (probably contextually ambiguous, possibly somewhat noisy) pairs of sentences and their meaning representations (logical forms).

Starting from these assumptions, we develop a Bayesian probabilistic account of semantically bootstrapped first-language acquisition in the child, based on techniques from computational parsing and interpretation of unrestricted text. Our learner jointly models (a) word learning: the mapping between components of the given sentential meaning and lexical words (or phrases) of the language, and (b) syntax learning: the projection of lexical elements onto sentences by universal construction-free syntactic rules. Using an incremental learning algorithm, we apply the model to a dataset of real syntactically complex child-directed utterances and (pseudo) logical forms, the latter including contextually plausible but irrelevant distractors. Taking the Eve section of the CHILDES corpus as input, the model simulates several well-documented phenomena from the developmental literature. In particular, the model exhibits syntactic bootstrapping effects (in which previously learned constructions facilitate the learning of novel words), sudden jumps in learning without explicit parameter setting, acceleration of word-learning (the "vocabulary spurt"), an initial bias favoring the learning of nouns over verbs, and one-shot learning of words and their meanings. The learner thus demonstrates how statistical learning over structured representations can provide a unified account for these seemingly disparate phenomena.

Keywords: language acquisition; syntactic bootstrapping; semantic bootstrapping; computational modeling; Bayesian model; cross-situational learning

[1] OA is now at the Departments of Computer Science & Cognitive Science, The Hebrew University of Jerusalem. TK is now at Google Research. NJS is now at the Berkeley Institute of Data Science, University of California, Berkeley.

1 Introduction

One of the fundamental challenges facing a child language learner is the problem of generalizing beyond the input. Using various social and other extralinguistic cues, a child may be able to work out the meaning of particular utterances they hear, like “you read the book” or “Eve will read *Lassie*”, if these are encountered in the appropriate contexts. But merely memorizing and reproducing earlier utterances is not enough: children must also somehow use these experiences to learn to produce and interpret novel utterances, like “you read *Lassie*” and “show me the book”. There are many proposals for how this might be achieved, but abstractly speaking it seems to require the ability to explicitly or implicitly (a) decompose the utterance’s form into syntactic units, (b) decompose the utterance’s meaning into semantic units, (c) learn lexical mappings between these syntactic and semantic units, and (d) learn the language-specific patterns that guide their recombination (so that e.g. “Eve will read *Lassie* to Fraser”, “will Eve read Fraser *Lassie*?”, and “will Fraser read Eve *Lassie*?” have different meanings, despite using the same or nearly the same words). A further challenge is that even in child-directed speech, many sentences are more complex than “you read *Lassie*”; the child’s input consists of a mixture of high- and low-frequency words falling into a variety of syntactic categories and arranged into a variety of more or less complex syntactic constructions.

In this work, we present a Bayesian language-learning model focused on the acquisition of *compositional* syntax and semantics in an *incremental, naturalistic setting*. That is, our model receives training examples consisting of whole utterances paired with noisy representations of the whole utterance’s meaning, and from these it learns probabilistic representations of the semantics and syntax of individual words, in such a way that it becomes able to recombine these words to understand novel utterances and express novel meanings. This requires that the model simultaneously learn how to parse syntactic constructions, assign meaning to specific words, and use syntactic regularities (for example, in verb argument structure) to guide interpretation of ambiguous input. Our training data consists of real, syntactically complex child-directed utterances drawn from a single child in the CHILDES corpus, and our training is incremental in the sense that the model is presented with each utterance exactly once, in the same order that the child actually encountered them.

The work described here represents an advance over previous models that focused on learning *either* word meanings *or* syntax given the other (see below for a review). By developing a joint learning model we are able to explore how these phenomena interact during learning. A handful of other joint learning models have been presented in the literature, but these have either worked from synthetic input with varying degrees of realism (Beekhuizen, 2015; Maurits, Perfors, & Navarro, 2009) or have not yet been evaluated on specific phenomena known from child language acquisition, as we do here (Chrupała, Kádár, & Alishahi, 2015; Jones, 2015). In particular, we show in a series of simulations that our model exhibits syntactic bootstrapping effects (in which previously learned constructions facilitate the learning of novel words), sudden jumps in learning without explicit parameter setting, acceleration of word-learning (the “vocabulary spurt”), an initial bias favoring the learning of nouns over verbs, and one-shot learning of words and their meanings. These results suggest that there is no need to postulate distinct learning mechanisms to explain these various phenomena; rather they can all be explained through a single mechanism of statistical learning over structured representations.

1.1 Theoretical underpinnings

Our model falls under the general umbrella of "Semantic Bootstrapping" theory, which assumes that the child can access a structural representation of the intended semantics or conceptual content of the utterance, and that such representations are sufficiently homomorphic to the syntax of the adult language for a mapping from sentences to meanings to be determined (Bowerman 1973; Brown 1973; E. Clark 1973; Grimshaw 1981; Pinker 1979; Schlesinger 1971; cf. Wexler and Culicover 1980:78-84; Berwick 1985:22-24). By "homomorphic", we simply mean that meaning representation and syntax stand in a "type-to-type" relation, according to which every syntactic type (such as the English intransitive verb) corresponds to a semantic type (such as the predicate), and every rule (such as English $\mathsf{S} \rightarrow \mathsf{NP}\ \ \mathsf{VP}$) corresponds to a semantic operation (such as function application of the predicate to the subject).

Early accounts of semantic bootstrapping (e.g. Wexler and Culicover 1980 and Berwick 1985) assumed perfect access to a single meaning representation in the form of an *Aspects*-style Deep Structure already aligned to the words of the language. Yet, as we shall see, semantic bootstrapping is sufficiently powerful that such strong assumptions are unnecessary.

Since, on the surface, languages differ in many ways—for example with respect to the order of heads and complements, and in whether such aspects of meaning as tense, causality, evidentiality, and information structure are explicitly marked—the meaning representations must be expressed in a universal prelinguistic conceptual representation, in whose terms all such distinctions are expressable. The mapping must further be learned by general principles that apply to all languages. These general principles are often referred to as "universal grammar", although the term is somewhat misleading in the present context since the model we develop is agnostic as to whether these principles are unique to language or apply more generally in cognition.

A number of specific instantiations of the semantic bootstrapping theory have been proposed over the years. For example, "parameter setting" accounts of language acquisition assume, following Chomsky (1981), that grammars for each natural language can be described by a finite number of finitely-valued parameters, such as head-position, pro-drop, or polysynthesis (Hyams, 1986, and much subsequent work). Language acquisition then takes a form that has been likened to a game of Twenty-Questions (Yang, 2006, Ch:7), whereby parameters can be set when the child encounters "triggers", or sentences that can only be analysed under one setting of a parameter. For example, for Hyams (1986), the fact that English has lexical expletive subjects (e.g., *it* in *it rained*) is unequivocal evidence that the pro-drop parameter is negative, while for others the position of the verb in simple intransitive sentences in Welsh is evidence for head-initiality. Such triggers are usually discussed in purely syntactic terms. However, in both examples, the child needs to know which of the words is the verb, which requires a prior stage of semantic bootstrapping at the level of the lexicon (Hyams 1986:132-133).

Unfortunately, parameter setting seems to raise as many questions as it answers. First, there are a number of uncertainties concerning the way the learner initially identifies the syntactic categories of the words, the specific inventory of parameters that are needed, and the aspects of the data that "trigger" their setting (Gibson and Wexler 1994; P. Niyogi and Berwick 1996; J. D. Fodor 1998). Second, several combinatoric problems arise from simplistic search strategies in this parameter space (J. D. Fodor & Sakas, 2005). Here, we will demonstrate that step-like learning curves used to argue for parameter-setting approaches (Thornton & Tesan, 2007) can be explained by a statistical model without explicit linguistic parameters.

A further variant of the semantic bootstrapping theory to be discussed below postulates a second, later, stage of "syntactic bootstrapping" (Braine, 1992; Gleitman, 1990; Landau & Gleitman, 1985; Trueswell & Gleitman, 2007), during which the existence of early semantically boostrapped syntax allows rapid or even "one-shot" learning of lexical items, including ones for which the situation of utterance offers little or no direct evidence. Early discussions of syntactic bootstrapping implied that it is a learning mechanism in its own right, distinct from semantic bootstrapping. However, we will demonstrate that these effects attributed to syntactic bootstrapping emerge naturally under the theory presented here. That is, our learner exhibits syntactic bootstrapping *effects* (using syntax to accelerate word learning) without the need for a distinct *mechanism*: the mechanism of semantic bootstrapping is sufficient to engender the effects.

Although varieties of semantic bootstrapping carry considerable currency, some researchers have pursued an alternative *distributional* approach (Redington, Chater, & Finch, 1998), which assumes that grammatical structure can be inferred from statistical properties of strings alone. Many proponents of this approach invoke Artificial Neural Network (ANN) computational models as an explanation for how this could be done—see Elman et al. (1996) for examples—while others in both cognitive science and computer science have proposed methods using structured probabilistic models (Cohn, Blunsom, & Goldwater, 2010; D. Klein & Manning, 2004; Perfors, Tenenbaum, & Regier, 2011). The distributional approach is appealing to some because it avoids the assumption that the child can access meanings expressed in a language of mind that is homomorphic to spoken language in the sense defined above, but inaccessible to adult introspection and whose detailed character is otherwise unknown.

There has been some success in using this kind of meaning-free approach to learn non-syntactic structure such as word- and syllable-level boundaries (Goldwater, Griffiths, & Johnson, 2009; Johnson & Goldwater, 2009; Phillips & Pearl, 2014). However, attempts to infer syntactic structures such as dependency or constituency structure, and even syntactic categories, have been notably less successful despite considerable effort (e.g. Abend, Reichart, & Rappoport, 2010; Christodoulopoulos, Goldwater, & Steedman, 2010; Cohn et al., 2010; D. Klein & Manning, 2004, 2005). Within the context of state-of-the-art natural language processing (NLP) applications, ANN models that have no explicit structured representations have yielded excellent language modeling performance (i.e., prediction of probable vs. improbable word sequences) (e.g., Mikolov, Karafiát, Burget, Cernockỳ, & Khudanpur, 2010; Sundermeyer, Schlüter, & Ney, 2012). They have also been used to learn distributed word representations that capture some important semantic and syntactic information (Mikolov, Yih, & Zweig, 2013). Yet the reason these models work as well as they do in NLP tasks arises from the way they mix sentence-internal syntax and semantics with pragmatics and frequency of collocation. Thus, they often learn to conflate antonyms as well as synonyms to similar representations (Turney & Pantel, 2010). This representation creates problems for the compositional semantics of logical operators (such as negation) of a kind that the child never exhibits.

1.2 Overview of the learner

In this work, we take a close relation between syntax and compositional semantics as a given. We also follow the basic premise of semantic bootstrapping that the learner is able to infer the meaning of at least some of the language she hears on the basis of nonlinguistic context. However, unlike some other versions of semantic bootstrapping, we assume that the available meanings are at the level of *utterances* rather than individual words, and that word meanings (i.e., the mapping from words to parts of the utterance meaning) are learned from such data.

We represent the meaning of an utterance as a *sentential logical form*. Thus, the aim of the learner is to generalize from the input pairs of an observed sentence and a possible meaning in order to interpret new sentences whose meaning is unavailable contextually, and to generate new sentences that express an intended meaning. Because of the limitations of available corpus annotations, the logical forms we consider are restricted to predicate-argument relations, and lack the interpersonal content whose importance in language acquisition is generally recognized. We examine the nature of such content in section 4.4, where we argue that our model can be expected to generalize to more realistic meaning representations.

Recent work suggests that the infant's physically limited view of the world, combined with social, gestural, prosodic, and other cues, may lead to considerably less ambiguity in inferring utterance meanings than has previously been supposed (C. Yu & Smith, 2012, 2013; Yurovsky, Smith, & Yu, 2013). Nevertheless, most contexts of utterance are likely to support more than one possible meaning, so it is likely that the child will have to cope with a number of distracting spurious meaning candidates. We model propositional ambiguity in the input available to the learner by assuming each input utterance s is paired with several contextually plausible logical forms $\{m_1,...,m_k\}$, of which only one is correct and the rest serve as distractors. While these meaning representations are assumed to reflect an internal language of mind, our model is general enough to allow the inclusion of various types of content in the logical forms, including social, information-structural, and perceptual.

Within this general framework, we develop a statistical learner that jointly models both (a) the mapping between components of the given sentential meaning and words (or phrases) of the language, and (b) the projection of lexical elements onto constituents and sentences by syntactic rules. In earlier work, we defined the learner and gave preliminary simulation results (Kwiatkowski, Goldwater, Zettlemoyer, & Steedman, 2012); here we expand considerably on the description of the learner, the range of simulations, and the discussion in relation to human language acquisition.

There has been considerable previous work by others on both word learning and syntactic acquisition, but they have until very recently been treated separately. Thus, models of cross-situational word learning have generally focused on learning either word-meaning mappings (mainly object referents) in the absence of syntax (Alishahi, Fazly, & Stevenson, 2008; M. Frank, Goodman, & Tenenbaum, 2009; McMurray, Horst, & Samuelson, 2012; Plunkett, Sinha, Møller, & Strandsby, 1992; Regier, 2005; Siskind, 1996; C. Yu & Ballard, 2007), or learning verb-argument structures assuming nouns and/or syntax are known (Alishahi & Stevenson, 2008, 2010; Barak, Fazly, & Stevenson, 2013; Beekhuizen, Bod, Fazly, Stevenson, & Verhagen, 2014; Chang, 2008; Morris, Cottrell, & Elman, 2000; S. Niyogi, 2002).[2] Conversely, most models of syntactic acquisition have considered learning from meaning-free word sequences alone (see discussion above), or have treated word-meaning mapping and syntactic learning as distinct stages of learning, with word meanings learned first followed by syntax (Buttery, 2006; Dominey & Boucher, 2005; Villavicencio, 2002).

[2] Most of these verb-learning models actually model argument structure *generalization* (e.g., predicting that if a verb has been seen in a double object construction, it might also be acceptable in a dative construction). We do not address this type of generalization directly in the model presented here. However, Kwiatkowski, Zettlemoyer, Goldwater, and Steedman (2011) have developed lexical generalization methods for similar models that could begin to address this problem.

Several previous researchers have demonstrated that correct (adult-like) knowledge of syntax (e.g., known part-of-speech categories or syntactic parses) can help with word-learning (Fazly, Alishahi, & Stevenson, 2010; Göksun, Küntay, & Naigles, 2008; Mellish, 1989; Thomforde & Steedman, 2011; Ural, Yuret, Ketrez, Koçbaş, & Küntay, 2009; H. Yu & Siskind, 2013), and a few (Alishahi & Chrupała, 2012; C. Yu, 2006) have gone further in showing that learned (and therefore imperfect) knowledge of POS categories can help with word learning. However, these models are not truly joint learners, since the learned semantics does not feed back into further refinement of POS categories.

By treating word learning and syntactic acquisition jointly, our proposal provides a working model of how these two aspects of language can be learned simultaneously in a mutually reinforcing way. And unlike models such as those of Maurits et al. (2009) and Beekhuizen (2015), our model learns from real corpus data, meaning it needs to handle variable-length sentences and predicates with differing numbers of arguments, as well as phenomena such as multiple predicates per sentence (including hierarchical relationships, e.g., *want to go*) and logical operators, such as negation and conjunction. To tackle this challenging scenario, we adopt techniques originally developed for the task of “semantic parsing" (more properly, semantic parser induction) in computational linguistics (Kwiatkowski, Zettlemoyer, Goldwater, & Steedman, 2010; Kwiatkowski et al., 2011; Thompson & Mooney, 2003; Zettlemoyer & Collins, 2005, 2007).

Our model rests on two key features which we believe to be critical to early syntactic and semantic acquisition in children. The first, shared by most of the models above, is statistical learning. Our model uses a probabilistic grammar and lexicon whose model parameters are updated using an incremental learning algorithm. (By parameters here, we mean the probabilities in the model; our model does not include linguistic parameters in the sense noted earlier of Chomsky (1981) and Hyams (1986).) This statistical framework allows the model to take advantage of incomplete information while being robust to noise. Although some statistical learners have been criticized for showing learning curves that are too gradual—unlike the sudden jumps in performance sometimes seen in children (Thornton & Tesan, 2007)—we show that our model does not suffer from this problem.

The second key feature of our model is its use of syntactically guided semantic compositionality. This concept lies at the heart of most linguistic theories, yet has rarely featured in previous computational models of acquisition. As noted above, many models have focused either on syntax or semantics alone, with another large group considering the syntax-semantics interface only as it applies to verb learning. Of those models that have considered both syntax and semantics for full sentences, many have assumed that the meaning of a sentence is simply the set of meanings of the words in that sentence (Alishahi & Chrupała, 2012; Allen & Seidenberg, 1999; Fazly et al., 2010; C. Yu, 2006). Connor, Fisher, and Roth (2012) addressed the acquisition of shallow semantic structures (predicate-argument structures and their semantic roles) from sentential meaning representations consisting of the set of semantic roles evoked by the sentence. Villavicencio (2002) and Buttery (2006) make similar assumptions to our own about semantic representation and composition, but also assume a separate stage of word learning prior to syntactic learning, with no flow of information from syntax back to word learning as in our joint model. Thus, their models are unable to capture syntactic bootstrapping effects. Chrupała et al. (2015) have a joint learning model, but no explicit syntactic or semantic structure. The model most similar to our own in this respect is that of Jones (2015), but as noted above, the simulations in that work are more limited than those we include here.

Much of the power of our model comes from this assumption that syntactic and semantic composition are closely coupled. To implement this assumption, we have based our model on Combinatory Categorial Grammar (CCG, Steedman, 1996b, 2000, 2012). CCG has been extensively applied to parsing and interpretation of unrestricted text (Auli & Lopez, 2011; S. Clark & Curran, 2004; Hockenmaier, 2003; Lewis & Steedman, 2014), and has received considerable attention recently in the computational literature on semantic parser induction (e.g., Artzi, Das, & Petrov, 2014; Krishnamurthy & Mitchell, 2014; Kwiatkowski et al., 2010, 2011; Matuszek, Fitzgerald, Zettlemoyer, Bo, & Fox, 2012; Zettlemoyer & Collins, 2005, 2007).

This attention stems from two essential properties of CCG. First, unlike Lexicalized Tree-Adjoining Grammar (Joshi & Schabes, 1997), Generalized Phrase Structure Grammar (Gazdar, Klein, Pullum, & Sag, 1985), and Head-driven PSG (Pollard & Sag, 1994)—but like LFG (Bresnan, 1982) and the Minimalist program (Chomsky, 1995)—a single nondisjunctive lexical entry governs both *in situ* and extracted arguments of the verb. Second, the latter kind of dependencies are established without the overhead for the learner of empty categories and functional uncertainty or movement, of the kind used in LFG and the Minimalist Program.

These properties of CCG, together with its low near-context-free expressive power and simplicity of expression, have made it attractive both for semantic parsing, and for our purposes here. Nevertheless, the presented framework can in principle be implemented with any compositional grammar formalism as long as (1) it allows for the effective enumeration of all possible syntactic/semantic derivations given a sentence paired with its meaning representation; and (2) it is associated with a probabilistic model that decomposes over these derivations.

Using our incremental learning algorithm, the learner is trained on utterances from the Eve corpus (Brown, 1973; Brown & Bellugi, 1964) in the CHILDES database (MacWhinney, 2000), with meaning representations produced automatically from an existing dependency annotation of the corpus (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010). We use these automatically produced meaning representations as a proxy for the child's actual meaning representation in the hidden conceptual language of mind. (Crucially, our model entirely ignores the alignment in these data between logical constants and English words, as if the sentences were in a completely unknown language.)

We evaluate our model in several ways. First, we test the learner's ability to correctly produce the meaning representations of sentences in the final Eve session (not included in the training data). This is a very harsh evaluation, since the learner is trained on only a small sample of the data the actual child Eve was exposed to by the relevant date. Nevertheless, we show a consistent increase in performance throughout learning and robustness to the presence of distractor meaning representations during training. Next, we perform simulations showing that a number of disparate phenomena from the language acquisition literature fall out naturally from our approach. These phenomena include sudden jumps in learning without explicit parameter setting; acceleration of word-learning (the "vocabulary spurt"); an initial bias favoring the learning of nouns over verbs; and one-shot learning of words and their meanings, including simulations of several experiments previously used to illustrate syntactic bootstrapping effects. The success of the model in replicating these findings argues that separate accounts of semantic and syntactic bootstrapping are unnecessary; rather, a joint learner employing statistical learning over structured representations provides a single unified account of both.

2 Semantic Bootstrapping for Grammar Acquisition

The premise of this work is that the child at the onset of language acquisition enjoys direct access to some form of pre-linguistic conceptual representations. We are not committed to these representations taking any particular symbolic or non-symbolic form, but for the compositional learning process to proceed, there must be some kind of structure in which complex concepts can be decomposed into more primitive concepts, and here we will abstractly represent this conceptual compositionality using a logical language. Such universal semantics or conceptual structure, broadly construed, has often been argued to be the most plausible source for the universal learning biases that enable language acquisition (Chomsky 1965:27-30; Crain and Nakayama 1987; 1995:54-55; Pinker 1979; Croft 2001; Ambridge, Pine, and Lieven 2014). The child's task is therefore to consider all the different ways that natural language allows chunks of meaning representation to be associated with lexical categories for the language they are presented with.

Concretely, we address the following learning problem. Given a corpus of transcribed utterances, paired with representations of their meaning (henceforth, their *logical forms*), the goal of learning is to induce a mapping between utterances (or transcribed text) to meaning representations, so as to support the correct interpretation of unseen text, namely by assigning it correct logical forms.

The model represents two types of information: a (probabilistic) pairing between lexical items and their meaning representation, and a distribution over the syntactic derivations the learner has been exposed to (a *probabilistic grammar*). The sensitivity of the infant learning language to statistical trends observed in the data has been demonstrated in multiple levels of linguistic representation (Gómez & Maye, 2005; Mintz, 2003; Saffran, Aslin, & Newport, 1996, *inter alia)* and is a key factor in this account of semantic bootstrapping.

Much of the difficulty in this setting stems from not being able to directly observe the meaning of individual words and the derivation trees, as the only supervision provided is of *sentential* meaning representations. In order to infer the probabilistic grammar and word meanings, the model entertains all possible derivations and word-meaning pairings, and weighs these based on their likelihood according to its hitherto acquired beliefs. This weighted space of derivations is subsequently used to update the beliefs of the learner when exposed to the next utterance.

2.1 Meaning Representations

Meanings of sentences in CCG (and in our model) are expressed in first-order predicate logic. The example

(1) a. "you like the doggies!"
 b. $\mathit{like}(\mathit{you}, \mathit{doggies})$

expresses a relation *like* between the entity *you* and the entity *doggies*.

Of course, we do not believe that the child's conceptual representation is as simple as this. As Tomasello (1999) has pointed out, the content that the child is actually working with and that actually provides it with the incentive to engage with language is highly interpersonal and social—probably more like something paraphraseable as "Mum is sharing my liking the doggies". Nevertheless, our model needs to represent the child's semantics somehow. For simplicity, we'll use terms of a naive English-centric lambda-calculus, and defer the question of what a more psychologically realistic human logical language might actually look like until section 4.4.

The lambda-calculus uses the λ-operator to define functions. These may be used to represent functional meanings of utterances but they may also be used as a “glue language” to compose elements of first order logical expressions like the above. For example, the function $\lambda x \lambda y.like(y,x)$ can be combined with the argument *doggies* to give the phrasal meaning $\lambda y.like(y, doggies)$ (corresponding to the VP “like the doggies”) via the lambda-calculus operation of *function application*. In function application, the formal argument x (introduced by the lambda notation) is replaced by *doggies* in the remainder of the expression.

Our model uses *typed* λ-expressions, where each variable and constant is given a semantic type. These types may either be atomic—such as e for entities (e.g., *you*, *doggies*) and t for truth values—or they may be complex—for instance, *sleep* is a function from entities to a truth value, of type (e,t), and *like* is a (Curried) function from pairs of entities to a truth value, of type $(e,(e,t))$. Lambda-calculus operations such as function application and composition are only allowed between forms of compatible types. For brevity, we will often not state semantic types explicitly.

Despite the shortcomings of such an approach to semantic representation, the logical forms we use do meet the minimal criteria of (1) not encoding word order information and (2) not encoding information about the correct segmentation of the utterance’s semantic forms above the level of atomic semantic symbols. For example, the logical form $like(you, doggies)$ does not determine the number of lexical items that will give rise to this composite representation. The space of possible derivations that can give rise to $like(you, doggies)$ includes the possibilities that there are four words in the sentence, one matching *like*, one matching *you* and one matching *doggies*, but also the possibility that the meaning is partitioned into other numbers of words, including the possibility that the whole meaning corresponds to a single word.

2.2 The Grammar

Combinatory Categorial Grammar (CCG) is a strongly lexicalised linguistic formalism that tightly couples syntactic types and corresponding semantic λ-terms. Each CCG lexical item in the lexicon L is a triplet, written

$$\text{word} \vdash \text{syntactic category} : \textit{semantic form}.$$

Examples include:[3]

$$\begin{aligned}
\text{you} &\vdash \mathsf{NP} : \textit{you} \\
\text{sleep} &\vdash \mathsf{S\backslash NP} : \lambda x.sleep(x) \\
\text{like} &\vdash \mathsf{(S\backslash NP)/NP} : \lambda x \lambda y.like(y,x) \\
\text{the} &\vdash \mathsf{NP/N} : \lambda f.the(x, f(x)) \\
\text{doggies} &\vdash \mathsf{N} : \lambda x.dogs(x)
\end{aligned}$$

Syntactic categories may be atomic (e.g., S or NP) or complex (e.g., $\mathsf{S\backslash NP}$ or $\mathsf{(S\backslash NP)/NP}$). Slash operators in complex categories define functions from the domain on the right of the slash to the result on the left in much the same way as lambda operators do in the corresponding semantic forms. The difference is that the direction of the slash operator defines the linear order of function

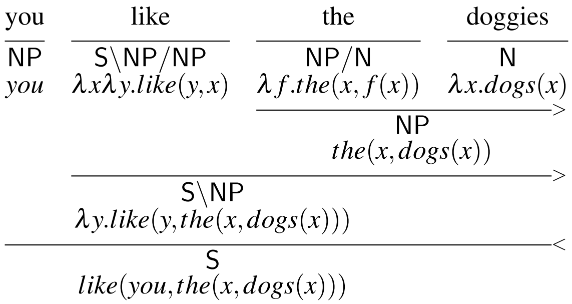
[3] Determiner semantics is represented as generalized quantifiers. Namely, $the(x, f(x))$ should be read as the generalized quantifier *the* over x, applied to $f(x)$.

and argument in the sentence. For example, $\mathsf{S/NP}$ corresponds to a constituent that along with an NP to its right forms a constituent with the category S, while $\mathsf{S\backslash NP}$ is similar but requires an NP to its left to form an S. For example, in the English sentence "you sleep", the verb "sleep" has the latter category, $\mathsf{S\backslash NP}$. In a verb-initial language like Welsh, the corresponding intransitive verb meaning "sleep" has the former category, $\mathsf{S/NP}$.

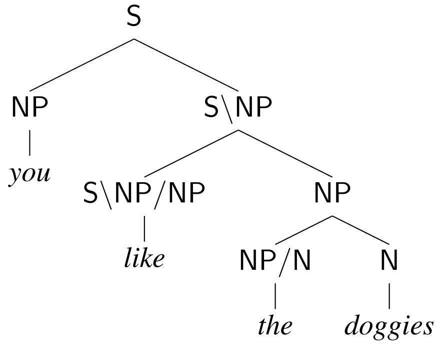
CCG uses a small set of *combinatory rules* to concurrently combine constituents according to their syntactic types and compose their semantic representations. Two simple combinatory rules are forward (>) and backward (<) *application*:

$$\begin{array}{llcll} \mathsf{X/Y}:f & \mathsf{Y}:g & \Rightarrow & \mathsf{X}:f(g) & (>) \\ \mathsf{Y}:g & \mathsf{X\backslash Y}:f & \Rightarrow & \mathsf{X}:f(g) & (<) \end{array}$$

Given the lexicon above, the phrase "you like the doggies" can be parsed using these rules as follows, where each step in the parse is labeled with the combinatory rule (> or <) that was used:



It is standard to write CCG parse trees as in the above diagram, with leaves at the top and combinatory operations written next to constituent demarcation lines. However, the same derivation could also be written in phrase-structure form as follows (semantics omitted):



A bit more needs to be said about transitive verbs. Unlike unary intransitive predicates and the determiner categories, transitive verbs in English, such as "like" in the example above, could in principle be associated with either of the two syntactic categories in (2), both of which support a derivation of the final logical form.

(2) a. like $\vdash \mathsf{(S\backslash NP)/NP} : \lambda x \lambda y.like(y,x)$
b. like $\vdash \mathsf{(S/NP)\backslash NP} : \lambda y \lambda x.like(y,x)$

However, we will assume here that the universally permitted set of transitive categories excludes cases of verb-medial categories (SVO or OVS) where, as in 2(b), the verb attaches first to the semantic subject.[4] The learner therefore considers only the following six possible categories/constituent orders for transitive verbs, only one of which will work for English:[5]

(3) a. SOV $\vdash \mathsf{(S\backslash NP)\backslash NP} : \lambda y \lambda x.\mathit{like}(x,y)$
b. SVO $\vdash \mathsf{(S\backslash NP)/NP} : \lambda y \lambda x.\mathit{like}(x,y)$
c. VSO $\vdash \mathsf{(S/NP)/NP} : \lambda x \lambda y.\mathit{like}(x,y)$
d. VOS $\vdash \mathsf{(S/NP)/NP} : \lambda y \lambda x.\mathit{like}(x,y)$
e. OVS $\vdash \mathsf{(S/NP)\backslash NP} : \lambda y \lambda x.\mathit{like}(x,y)$
f. OSV $\vdash \mathsf{(S\backslash NP)\backslash NP} : \lambda x \lambda y.\mathit{like}(x,y)$

Finally, we note that our model also includes CCG combinatory rules of forward ($> \mathbf{B}$) and backward ($< \mathbf{B}$) *composition*:

$$\begin{array}{llcll} \mathsf{X/Y} : f & \mathsf{Y/Z} : g & \Rightarrow & \mathsf{X/Z} : \lambda x.f(g(x)) & (> \mathbf{B}) \\ \mathsf{Y\backslash Z} : g & \mathsf{X\backslash Y} : f & \Rightarrow & \mathsf{X\backslash Z} : \lambda x.f(g(x)) & (< \mathbf{B}) \end{array}$$

These rules are crucial to the CCG analysis of unbounded dependency-inducing constructions such as relativization, to which we return in the discussion of later learning in section 4.4.

2.3 The Probabilistic Model

The goal of the presented model is to learn a form-meaning mapping that can support the correct interpretation of unseen text. We assume that this mapping involves a latent syntactic structure that mediates between phonological/orthographic and semantic form. Given the uncertainty involved in grammar induction (as a finite number of examples does not uniquely determine a CCG grammar, or any other plausible class of grammars), we formulate the problem of acquisition as one of probabilistic inference.

There are two ways to construe the use of probabilistic modeling. In one, probabilities serve to make rational inferences in an uncertain environment. In this case, the target grammar can be formulated without the use of probabilities, while the probabilistic model represents the "belief" of the model in each of the possible non-probabilistic hypotheses. Alternatively, the target grammar itself is assumed to be probabilistic, either because it captures information that is probabilistic in nature or because the model does not have access to some of the information that would be needed to disambiguate (for example, semantic distinctions not represented in the logical forms we use). Our model and results are consistent with either interpretation; we will return to discuss this point in Section 4.

We take a Bayesian inference approach and explore learning through an idealized rational learner that induces the grammar most supported by the observed data given the assumptions made by the model. We depart from this idealized setting only in assuming that learning is done incrementally, so that the learner has no access to previously observed examples beyond what has already been learned from them (i.e., the updated parameters of the statistical model). Using an incremental

[4] This assumption rules out verb-medial ergative languages but can be circumvented by subcategorizing the subject and object NPs based on their case.

[5] We assume, following Baldridge (2002), that free word-order languages simply have more than one of these categories.

algorithm makes it easy to examine how the learner's grammar and behavior develop over time as examples are presented, and is of course more psychologically plausible than a batch algorithm.

As discussed above, our model takes as input a sequence of sentences, each paired with one or more possible meanings. For the moment, assume each sentence s is paired with a single meaning m. The learner also maintains a probability distribution over model parameters Θ. These parameters specify the probabilities of using or combining lexical items into phrases according to different rules, and can be thought of as the learner's grammar (including the lexicon, as CCG is a lexicalized formalism). The distribution over Θ thus represents the learner's belief about which grammars are most plausible given the input so far.

Before observing a particular (s,m) pair, the learner has some *prior* probability distribution over the grammar, $P(\Theta)$. At the beginning of the learning process, this prior will be entirely flat, and will assign equal probability to symmetric options (e.g., verb-final vs. verb-initial) as the learner doesn't know anything specific about the language being learned.[6] However, the prior will be updated with each new example. Upon observing an (s,m) pair, the learner infers the probabilities of all possible syntactic derivation trees t that could provide a derivation of the (s,m) pair, along with a *posterior* distribution over Θ. That is, the learner infers a distribution $P(t,\Theta|s,m)$, which can be expressed using Bayes' rule as:

$$P(t,\Theta \mid s,m) = \frac{P(s,m,t \mid \Theta)P(\Theta)}{P(s,m)} \tag{1}$$

As learning progresses, the prior for each new example is in fact the posterior distribution learned after observing the previous example, so over time the prior will become much more peaked, assigning high probability to lexical items and rules used frequently in the previously inferred derivations.

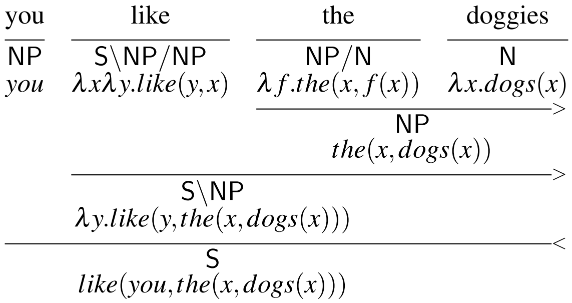
For the model to be complete, we also need to specify how to compute the *likelihood* $P(s,m,t|\Theta)$, which determines how probable different derivations are under a particular grammar. (For any (s,m) pair, $P(s,m)$ is a constant that simply normalizes the posterior distribution to sum to 1, so we don't need to define it explicitly.)

As is usual in Bayesian models, we will define the likelihood using a generative formulation, i.e., we will define a probabilistic process that describes how (s,m,t) triples (derivations) are assumed to be generated given a particular set of parameters Θ. This process will consist of multiple steps; by multiplying together the probabilities at each step we can determine the probability of the final complete derivation $P(s,m,t \mid \Theta)$. The learner that we present in Section 2.4 incrementally changes the parameters Θ to increase the probabilities of derivations that support the (s,m) pairs seen during training.[7]

Consider now a particular (s,m,t) derivation, such as the one given above for "you like the doggies" : $\mathit{like}(\mathit{you},\mathit{the}(x,\mathit{dogs}(x)))$ and repeated here:

[6] There is nothing in our model to exclude asymmetric priors, such as those proposed by Culbertson, Smolensky, and Wilson (2013). We assume a flat initial prior as it is the most conservative option.

[7] Since the syntactic tree t for any particular (s,m) pair is never directly observed, the likelihood term favors values of Θ that assign high probability to the *set of possible derivations* $t_{(s,m)}$ of the (s,m) pair in question—Appendix B describes how the space of all possible derivations is defined. When multiple (s,m) pairs are observed, the likelihood favors Θ that assign high probability to the set of all derivations of all observed pairs. This goal can be achieved by assigning high probability to derivations with subparts that can be used to explain many different (s,m) pairs. Assigning high probability to an idiosyncratic parse could raise the probability of $t_{(s,m)}$ for a particular (s,m) pair, but will steal probability mass from other parses that are more generally applicable, thus making the possible derivations for all other (s,m) pairs less likely.



The probability of such a derivation will have three parts. First, the probability of all of the function applications needed to decompose each parent node into each child combination of nodes. These are essentially the rules of the syntax. Second, the probability with which each of the resulting leaf syntactic category child nodes (e.g., NP or NP/N) is assigned a given meaning (e.g., *you* or $\lambda f.the(x,f(x))$). Third, the probability with which each meaning corresponds to a word in the sentence (e.g., 'you' or 'the').

Formally, let $w_1,...,w_n$ be the words in the sentence s, $m_1,...,m_n$ be their meaning representations (i.e., the meanings of each leaf in the derivation tree), and $c_1,...,c_n$ be their syntactic categories. The derivation is generated as follows:

1. Generate each syntactic rule in the derivation, with probability $P(\gamma \mid a)$ for a rule $a \to \gamma$ with parent category a and child categories γ. Each parent category X either generates a pair of child categories through inverse application or composition (e.g., X may generate X/Y and Y; X/Z may generate X/Y and Y/Z; See Appendix B for the formal definition of this procedure), or generates a leaf by generating the symbol $\mathsf{X}_{\mathsf{LEX}}$. We assume all derivations start with the START symbol, but for brevity omit this and the $\mathsf{X}_{\mathsf{LEX}}$ categories from the example derivations shown. The probability of this step is given as:

(2) $$P_{SYNTAX}(t) = \prod_{a \to \gamma \in t} P(\gamma \mid a)$$

2. For each leaf syntactic category c_i in $c_1,...,c_n$, generate the corresponding meaning representation m_i with probability $P(m_i \mid c_i)$. The probability of this step is given as:

(3) $$P_{MEANINGS}(m|t) = \prod_{i=1}^{n} P(m_i \mid c_i)$$

3. For each m_i, generate the corresponding word w_i with probability $P(w_i \mid m_i)$. The probability of this step is given as:

(4) $$P_{WORDS}(s|m) = \prod_{i=1}^{n} P(w_i \mid m_i)$$

Putting these steps together, we get the overall probability of a derivation:

$$P(s,m,t) = P_{SYNTAX}(t) \cdot P_{MEANINGS}(m|t) \cdot P_{WORDS}(s|m) \tag{5}$$

For instance, the probability of the above derivation of “you like the doggies” is the product of the following three terms:

$$\begin{aligned}
P_{SYNTAX}(t) =& P(\mathsf{S} \mid \mathsf{START}) \times P(\mathsf{NP},\ \mathsf{S\backslash NP} \mid \mathsf{S}) \times P(\mathsf{S\backslash NP/NP},\ \mathsf{S\backslash NP} \mid \mathsf{S\backslash NP}) \times \\
& P(\mathsf{NP/N},\ \mathsf{N} \mid \mathsf{NP}) \times P(\mathsf{NP/N},\ \mathsf{N} \mid \mathsf{NP}) \times \\
& P(\mathsf{NP_{LEX}} \mid \mathsf{NP}) \times P([\mathsf{S\backslash NP/NP}]_{\mathsf{LEX}} \mid \mathsf{S\backslash NP/NP}) \times \\
& P([\mathsf{NP/N}]_{\mathsf{LEX}} \mid \mathsf{NP/N}) \times P(\mathsf{N_{LEX}} \mid \mathsf{N}) \\
P_{MEANINGS}(m|t) =& P(you \mid \mathsf{NP_{LEX}}) \times P(\lambda x \lambda y.like(y,x) \mid [\mathsf{S\backslash NP/NP}]_{\mathsf{LEX}}) \times \\
& P(\lambda f.the(x, f(x)) \mid [\mathsf{NP/N}]_{\mathsf{LEX}}) \times P(\lambda x.dogs(x) \mid \mathsf{N_{LEX}}) \\
P_{WORDS}(s|m) =& P(\text{you} \mid you) \times P(\text{like} \mid \lambda x \lambda y.like(y,x)) \times \\
& P(\text{the} \mid \lambda f.the(x, f(x))) \times P(\text{dogs} \mid \lambda x.dogs(x))
\end{aligned}$$

This model is essentially a probabilistic context-free grammar (PCFG). While naive PCFGs impose conditional independence assumptions that are too strong to characterize natural language, they can be augmented to overcome this problem (Charniak, 1997; Collins, 1997). PCFGs are a simple and well-understood class of models with standard parsing and learning algorithms, and are powerful enough to do well in prediction of both phrase-structure trees (Collins, 1997) and CCG derivation trees (Hockenmaier & Steedman, 2002) in naturalistic settings. Future work will explore the use of richer families of distributions.

In order to complete the model, we still need to define how the distributions used in Equations 2, 3 and 4 are calculated: $P(\gamma \mid a)$, $P(m \mid c)$, and $P(w \mid m)$. First, consider $P(\gamma \mid a)$, the distribution over possible expansions γ given a syntactic production head a. Notice that this distribution needs to be defined over an infinite set of possible expansions, since the learner does not know ahead of time which expansions will be needed for their language. Recent probabilistic models of language acquisition (Alishahi & Stevenson, 2008, 2010; Feldman, Griffiths, Goldwater, & Morgan, 2013; S. Frank, Feldman, & Goldwater, 2014; Goldwater et al., 2009), have demonstrated that Dirichlet Processes can be used to successfully model such infinite distributions, so we adopt that approach here. When applied to our problem (see Appendix A for details), the Dirichlet Process defines the probability of a particular expansion given the production head as

$$P(\gamma \mid a) = \frac{n_{a\to\gamma} + \alpha_{syn} H_a}{n_a + \alpha_{syn}} \tag{6}$$

where $n_{a\to\gamma}$ is the number of times the rule $a \to \gamma$ has been used previously in a derivation, n_a is the number of times any rule expanding a has been used, and α_{syn} is a fixed parameter of the model which determines the learner’s tendency to infer derivations containing (any) new rule that has not been used previously. H_a, another fixed parameter of the model, is a prior distribution over rules: of all the rules headed by a that have not been used before, which ones are more likely to be used?

Notice that under this definition, the more a rule has been used relative to other expansions of the same parent, the higher its probability will be. Over time, rules with higher probability will be used more, incurring higher counts, which raises their probability even more: a virtuous cycle of learning. However, even rules that have never been used before (with $n_{a\to\gamma}$ =0) have some non-zero probability associated with them, as determined by the values of α_{syn} and H_a. We define H_a without any particular bias towards any of the CCG operators, or any of the atomic syntactic categories. However, we do include a bias towards simpler syntactic categories (e.g., expanding an NP into $(\mathsf{NP/N},\mathsf{N})$ would be *a priori* more likely than expanding it into $(\mathsf{NP/(N/N)},\mathsf{N/N}))$.

The remaining two distributions $P(m \mid c)$ (the distribution over leaf meaning representations given the syntactic category) and $P(w \mid m)$ (the distribution over wordforms given a meaning representation m) are also defined using Dirichlet Processes. Again, these provide well-defined distributions over the infinite sets of possible meaning representations and wordforms, assigning higher probability to meanings and words that have been seen more frequently, but non-zero probability to novel meanings and words. The prior distribution over unseen meanings prefers simpler logical expressions (those containing fewer logical constants), and the prior distribution over unseen words prefers shorter words. See Appendix A for details.

2.4 The Learning Algorithm

The previous subsections described the input, internal representations, and probabilistic model used by our learner. This subsection explains how the learner processes the input data in order to determine the possible derivations of each (sentence, meaning) pair and update its knowledge of the parameters Θ on the basis of the input and its derivations.

At a high level, the learning algorithm is straightforward. The corpus contains sentences $s_1 \ldots s_N$, and we assume for the moment that each s_i is paired with a single unambiguous meaning representation m_i. The learner processes the corpus incrementally, at each step considering the pair (s_i, m_i) and computing the set of possible derivations $\mathbf{t}_i$ for this pair. Roughly speaking, a single derivation can be found by recursively splitting s_i and m_i into smaller and smaller chunks that, when recombined using the CCG application and composition rules, will yield the original input pair. By considering all possible splits at each step in this process, all possible derivations can be found and their probabilities computed based on the learner's current beliefs about Θ and the probabilistic model described in Section 2.3.[8] Details of the splitting process, including how the possible syntactic categories are determined for each chunk, can be found in Appendix B.

Next, the learner updates its beliefs about Θ. Each derivation in $\mathbf{t}_i$ will include binary (combinatory) rules as well as unary rules that generate the leaf syntactic categories, meaning representations, and wordforms. The learner updates its probabilities for any rules used in $\mathbf{t}_i$, weighting the number of times a rule is used by the probability of the derivation(s) in which it is used. In other words, the learner will increase the probability of a rule more if it is used frequently, or in a derivation that has high probability according to the current model parameters. It is also important to account for the fact that early on, the learner's prior is based on little evidence, whereas later it is based on much more evidence. We therefore assume that early in learning the learner makes larger updates, weighting the evidence from the current example more heavily relative to the model's previously learned prior. The rate at which learning drops off is controlled by an additional fixed parameter, the *learning rate*. A more gradual change in the learning rate means the learner will continue to make large updates for longer, effectively forgetting more about early examples and

[8] We efficiently store and compute probabilities over the set of possible derivations by using a "packed forest" representation of the kind used in chart parsing in computational linguistics.

taking longer to converge. Details of the update procedure can be found in Appendix C.[9]

The only change needed in order to handle the more realistic condition where the child/learner has to consider more than one possible meaning for each sentence is that the set of derivations considered is the union of all possible derivations for the current sentence paired with each possible meaning. In this scenario, rules that are used in derivations for the incorrect meaning may gain some probability mass, especially early on when the learner has little evidence of the correct rules. However, they will quickly be swamped by the correct rules, for the same reason that rules used in incorrect derivations of the correct meaning are. That is, the correct rules apply consistently (and therefore frequently) across a wide range of sentences, leading to increasingly high probabilities for those rules and a virtuous cycle of learning. As we shall see, even when there is an unknown word in the sentence, the learner's increasing knowledge about the rules that are typically used in other sentences can lead to very rapid learning of new words (both their meanings and syntax) later on in learning.

To illustrate the learning algorithm, consider a very simple example in which the child's first input is "more doggies", which we will assume, simplifying as usual, that she understands as *more*(*dogs*). Upon hearing this sentence, the child inversely applies the CCG combinators to retrieve every derivation possible under CCG that yields the input pair "more doggies" : *more*(*dogs*). In this case, there are three possible derivations:

(4) a.
$$\dfrac{\dfrac{\text{more}}{\mathsf{NP/N} : \mathit{more}} \quad \dfrac{\text{doggies}}{\mathsf{N} : \mathit{dogs}}}{\mathsf{NP} : \mathit{more}(\mathit{dogs})}{>}$$

b.
$$\dfrac{\dfrac{\text{more}}{\mathsf{N} : \mathit{dogs}} \quad \dfrac{\text{doggies}}{\mathsf{NP\backslash N} : \mathit{more}}}{\mathsf{NP} : \mathit{more}(\mathit{dogs})}{<}$$

c.
$$\dfrac{\text{more doggies}}{\mathsf{NP} : \mathit{more}(\mathit{dogs})}$$

The following set of candidate lexical entries can be read off the three derivations in (4):

(5) a. more $\vdash \mathsf{NP/N} : \mathit{more}$
b. more $\vdash \mathsf{N} : \mathit{dogs}$
c. doggies $\vdash \mathsf{NP/N} : \mathit{more}$
d. doggies $\vdash \mathsf{N} : \mathit{dogs}$
e. more doggies $\vdash \mathsf{NP} : \mathit{more}(\mathit{dogs})$

Since this is the first utterance the child hears, the correct lexical entries (a, d) and incorrect entries (b, c) are completely symmetrical and therefore equiprobable: if the child wants to say *more* she is as likely to choose "dogs" as "more". Lexical entry (e) will have lower probability because the priors over wordforms and logical forms favor shorter and simpler entries.

Now suppose the child encounters a second utterance, "more cookies", which will have three

[9] We show later that within broad limits, the learning rate has no interesting effects on the overall performance of the learner.

derivations and five possible lexical entries analogous to those for “more doggies”, with the correct derivation being:

(6)
$$\begin{array}{cc} \text{more} & \text{cookies} \\ \hline \mathsf{NP/N} : \mathit{more} & \mathsf{N} : \mathit{cookies} \\ \hline \multicolumn{2}{c}{\mathsf{NP} : \mathit{more}(\mathit{cookies})} \end{array} >$$

In this case the correct derivation will receive higher probability than the incorrect ones, due to the child’s previous experience with “more doggies”. In particular, because “more” : *more* has occurred already in a possible derivation, $P(\text{more} \mid \mathit{more})$ will be higher than $P(\text{more} \mid \mathit{cookies})$ or $P(\text{cookies} \mid \mathit{more})$, and this will give a higher probability overall to the correct derivation, even though at this point there is no preference based on syntax. That is, prior to observing “more cookies”, noun-initial and noun-final NPs have both been seen once in equiprobable derivations, so $P(\mathsf{N},\ \mathsf{NP/N} \mid \mathsf{NP})$ is equal to $P(\mathsf{NP\backslash N},\ \mathsf{N} \mid \mathsf{NP})$. Importantly for future learning, however, the child is now able to update her syntactic probabilities based on this example. Because the derivation in (6) has higher probability than the competing derivations, the $\mathsf{NP} \rightarrow \mathsf{NP/N}\ \ \mathsf{N}$ rule will be updated to receive higher probability than the alternative $\mathsf{NP} \rightarrow \mathsf{N}\ \ \mathsf{NP\backslash N}$, and in future examples this will allow the child to use syntax to help infer the meanings of novel words. Our simulations in Sections 3.4 and 3.5 suggest that this effect leads to an acceleration of learning over time, and Section 3.7 shows that the same effect can explain results from classic syntactic bootstrapping experiments, where children use word order to correctly learn the meaning of a novel verb used with two familiar nouns (e.g., *the man daxed the baby*).

It should be noted that, due to the incremental nature of our learning algorithm, observing an utterance like “more cookies” does not cause the model probabilities for *dogs* to be updated; i.e., the probability that *dogs* is realized as “doggies” will still be equal to the probability that it is realized as “more”. This behaviour would differ with a batch learning algorithm, since “more cookies” provides evidence that *more* is realized as “more”, so that in “more doggies”, *dogs* is more likely providing the semantics for “doggies”. In our incremental algorithm, the parameters for any particular word are only updated when that word is seen. In that sense, the model built by the incremental learner is an approximation to what would be learned by an exact batch learner.

2.5 Generating meanings for novel sentences

To evaluate our account of language learning, we apply the resulting parser to unseen sentences and ask it to supply the corresponding meaning representation. Because of the generative formulation of the model, there is a straightforward probabilistic interpretation of this task. In particular, given an input sentence s and the model’s current estimate of the parameters Θ, we predict the most probable, i.e., *maximum a posteriori* (MAP), meaning representation m^* and syntactic tree t^*:

$$\begin{aligned} &(7) & t^*, m^* &= \operatorname*{argmax}_{m,t} P(m,t \mid s,\Theta) \\ &(8) & &= \operatorname*{argmax}_{m,t} \frac{P(m,t,s \mid \Theta)}{P(s \mid \Theta)} \\ &(9) & &= \operatorname*{argmax}_{m,t} P(m,t,s \mid \Theta) \end{aligned}$$

where Eq (9) follows from (8) because $P(s \mid \Theta)$ doesn’t depend on m or t.

As the model is an instance of a PCFG, we use the standard probabilistic CKY algorithm for computing the most likely (m,t,s) triplet given the parameters of the model.[10] We note that the model is unable to predict m if s contains unknown words, as the logical constants used to represent them can be any of an infinite set of unknown items. Instead, the model predicts the most likely *shell* logical form, where anonymous but semantically typed place-holders replace unseen constants. For example, suppose the model sees the sentence "Mary blicks John" with no paired logical form (corresponding to a situation where the child hears the sentence but cannot infer the semantics from non-linguistic context). If it has learned enough about the syntax and semantics of other verbs and about the meaning of "Mary" and "John", then it will be able to infer the meaning of the sentence as $PLACE\_HOLDER(Mary,John)$, indicating that it knows "blick" is a verb taking two arguments in a particular order, but does not know the specific lexical meaning of "blick". We also investigate learning of this kind, using artificial unseen words.

3 Simulations

We conduct a range of simulations with our model, looking at four main types of effects: (1) learning curves, both in terms of the model's overall generalization ability, and its learning of specific grammatical phenomena; (2) syntactic bootstrapping effects, where previously acquired constructions accelerate the pace at which words are learned (we show overall trends and simulate specific behavioral experiments); (3) one-shot learning effects, showing that the model is able to infer the meaning of newly observed words, given that enough about the syntax and the rest of the words in the sentence is known; (4) findings as to the relative pace of learning of nouns and verbs.

Our simulations are summarized in Table 1 and described in detail below, following the description of the input data.

3.1 Input to the Learner

3.1.1 Corpus. We know of no corpus of child-directed utterances manually annotated with compatible logical forms. Instead, we use the Eve corpus (Brown & Bellugi, 1964) in the CHILDES database (MacWhinney, 2000) and construct the logical representations semi-automatically based on the dependency and morphosyntactic category annotations of Sagae et al. (2010), using the general method of Çakıcı (2005) and Ambati, Deoskar, and Steedman (2016).

The Eve corpus contains 20 roughly two-hour sessions, two per month, collected longitudinally from a single child aged 18-27 months. The dependency annotations were converted to logical forms using a manually defined mapping that addresses major constructions, such as verb-argument structure, coordination, WH and yes/no questions, copulas and auxiliaries. The mapping covers 41% of the utterances in the corpus, yielding 5831 utterances. About a quarter of the discarded utterances consist of one-word interjections such as "hmm" or "yeah". We further filtered out sentences of 10 words or more, and of 10 or more extractable sub-expressions, leaving a final corpus of 5123 utterances. We divide this into a training set of 4915 utterances (sessions 1–19), and a held-out test set of 208 utterances (session 20). Some example sentence-meaning pairs can be found in Appendix D[11].

[10] Strictly speaking, the model is an *infinite PCFG*, which (unlike a standard PCFG), places some probability mass on unseen expansions. However, such expansions are excluded when predicting the MAP representation.

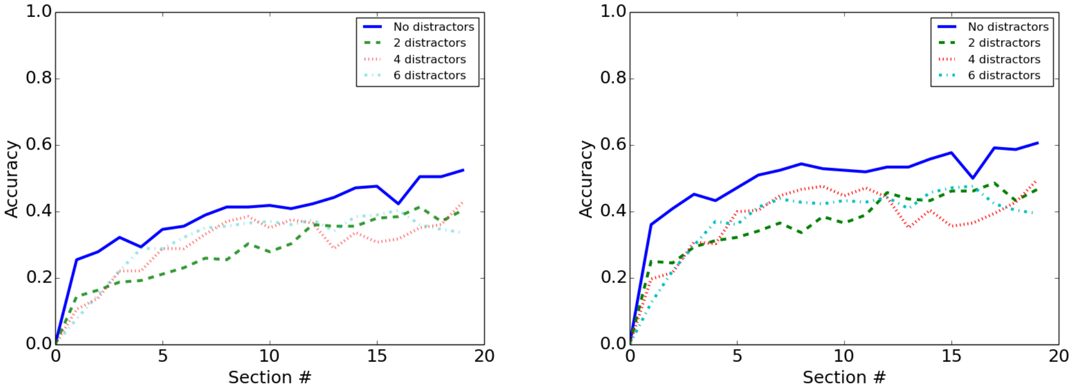
[11] The derived corpus is available online through the corresponding author's website.

Section	Figure(s)	Simulation and Main Effect
3.2	1	Parsing accuracy improves over time; performance is robust to propositional uncertainty.
3.3	2, 3, 4, 5	The model learns word ordering: that transitive sentences are SVO, and that determiners and prepositions are pre-nominal. Steep learning curves are exhibited despite having no explicit parameter-setting.
3.4	6, 7, 8, 9	The model demonstrates one-shot learning of verbs, nouns, determiners and prepositions in later stages of acquisition. These results can be seen as a naturalistic (corpus-based) analogue to the carefully controlled simulations reported in Figures 14-16, and therefore as a separate illustration of how learned syntactic knowledge can help accelerate word learning (the "syntactic bootstrapping" effect; Gleitman 1990).
3.5	10, 11	The model exhibits acceleration in the learning pace of transitive verbs and of nouns, similar to the "vocabulary spurt" observed in children (Reznick & Goldfield, 1992). Again, these results can be viewed as (partly) due to syntactic bootstrapping.
3.6	12, 13	The model obtains a larger production vocabulary of nouns than of verbs, despite not being explicitly biased to do so (Gentner, 1982; Tomasello, 1992:210). Moreover, in the presence of distractors, verbs are initially under-represented in the learner's production vocabulary (compared to nouns), relative to what can be predicted from the verb to noun input ratio in the child-directed utterances (Gentner & Boroditsky, 2001 and references therein).
3.7	14, 15, 16	The model simulates results from classic syntactic bootstrapping experiments. It performs one-shot learning, using its acquired knowledge of syntax to disambiguate between possible meanings for a transitive verb (following Fisher, Hall, Rakowitz, and Gleitman 1994) and between noun-like and preposition-like meanings for a novel word (following Fisher, Klingler, and Song 2006).

Table 1
Navigation table for the presented simulations and main effects.

This is the only dataset of its kind that we know of. However, it should be noted that it constitutes what we estimate to be less than 2% of the data that the actual child Eve was learning from over this same period in order to achieve the competence implicit in session 20, the held-out test set. It will become apparent below that this dataset is only just large enough to make stable and convergent learning possible, given our current model and learning algorithm.

3.1.2 Propositional Uncertainty. In order to relax the overly optimistic assumption that the child is exposed to entirely unambiguous situations, we add some uncertainty as to the correct logical form. We do so by pairing each utterance with several logical forms, only one of the which is the correct one. Rather than choosing distractor logical forms at random, we follow Fazly et al. (2010) in choosing distractors from nearby utterances in the corpus, so they are likely to share some semantic content with correct logical forms (as they were uttered in a similar situation). In particular, distractors are taken to be the logical forms of the w utterances preceding and following the target training example, where $w \in \{0,1,2,3\}$. We hence explore four settings: one with no propositional uncertainty (target logical form alone), and three with increasing amount of propositional uncertainty (target plus two, four and six distractors).



(a) (b)

Figure 1. The model is trained by incrementally viewing sentences from the first i (x-axis) sessions, each paired with a correct meaning and up to six distractor meanings; it is then tested on its accuracy (y-axis) at correctly assigning full sentence-level meanings to unseen test set sentences, using the 20th session as a test set. In Figure 1(a), sentences containing any previously unseen word are automatically marked as incorrect (since no specific semantics can be assigned to that word); in Figure 1(b), the model can guess placeholder meanings for unseen words, and receives credit if the rest of the semantic form is correct.

3.2 Parsing Unseen Sentences

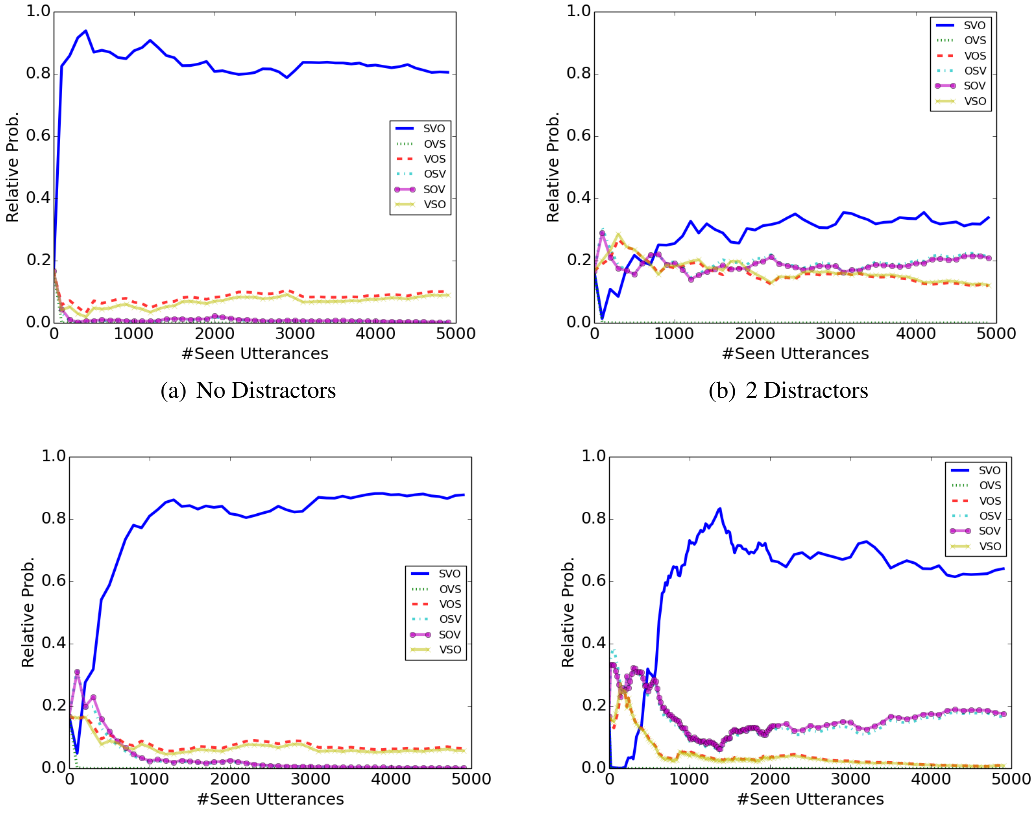
We begin by evaluating our parser's learning performance by training it on varying amounts of data, from 1-19 sessions. After training on the first i sessions (for $i = 1 \ldots 19$), we test the learner's ability to parse the sentences in the held-out test file (session 20). We use a single test file rather than testing each time on session $i+1$ to isolate the changes in performance due to learning, rather than confounding results with possible changes in the test data over time.

For each utterance in the test file, we use the parsing model to predict the MAP estimate for m^* and compare this to the target meaning. As described in Section 2.5, when a word has never been seen at training time our parser has the ability to "guess" a shell logical form with placeholders for unseen constant and predicate names. Figure 1 presents the prediction accuracy, i.e., the proportion of utterances for which the prediction m^* matches the target meaning with and without word-meaning guessing.

The low performance at all levels of uncertainty is largely due to the sparsity of the data with 21% of all training sentences containing a previously unseen word. Nevertheless, while increasing the propositional uncertainty slows the learning down, the learner shows increasing performance over time in all cases. This is an indication of the model's robustness to noise, a major issue of concern in computational models of acquisition (Siskind, 1996; Yang, 2002).

3.3 Learning word order

We next examine the model's learning of syntax, starting with transitive verbs. There are eight possible "word orders" (categories including semantic forms) that a transitive may be associated with, two of which are only applicable to ergative languages, and are ruled out by stipulation (see Section 2.2). The proposed model does not have a parameter that directly represents its belief as



(a) No Distractors

(b) 2 Distractors

(c) 4 Distractors

(d) 6 Distractors

Figure 2. Learning that English is an SVO language. Plots show the relative posterior probability assigned by the model to the six possible categories of transitive verbs. The x-axis is the number of training utterances seen so far. In the 2 distractor setting, the failure to assign most of the probability mass to the correct SVO category is an artefact of the small dataset (see text).

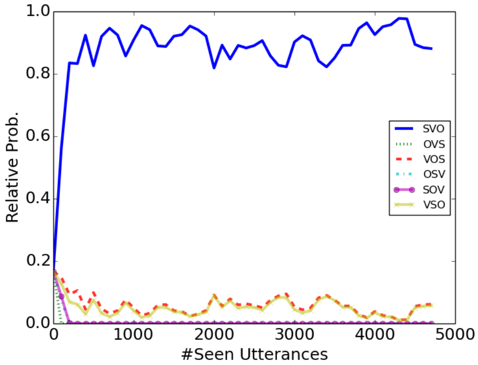


Figure 3. The relative probability of the six categories considered for transitive verbs in the 2 distractor setting, as a function of the number of seen utterances, where the learning rate is modified to be more gradual. The correct SVO category is learned quickly and effectively.

to which of the categories is the correct one. However, the model does maintain a distribution over syntactic expansions and logical forms, which we use to compute the model's learned prior distribution over the six remaining transitive verb categories (where the model's beliefs after seeing the training data are its prior over the categories used to parse any new utterance).

Figure 2 presents the relative probability for the transitive verb categories in the various settings. In all cases, the correct SVO category is learned to be the most probable upon convergence. In the three cases with propositional uncertainty, we see that the SVO category receives very little probability after 100 iterations (after an initial probability of $\frac{1}{6}$). The model is, however, able to recover, and eventually the correct SVO order is learned to be the most probable. However, in the 2 distractor setting the SVO category is never overwhelmingly favored over the other categories (as it is in all the other settings). This stems from the difficulty in setting a learning rate that works well across many different experimental settings given the small size of the training set. Consequently, in this case the algorithm does not manage to recover from its initial under-estimation of the relative probability of SVO, since the learning rate is too steep to allow effectively weighing in further evidence observed later in learning. To confirm this, we experimented with a more gradual learning rate (details in Appendix C), and present the relative probabilities of the six categories for this setting in Figure 3, which indeed shows a clear preference for SVO.

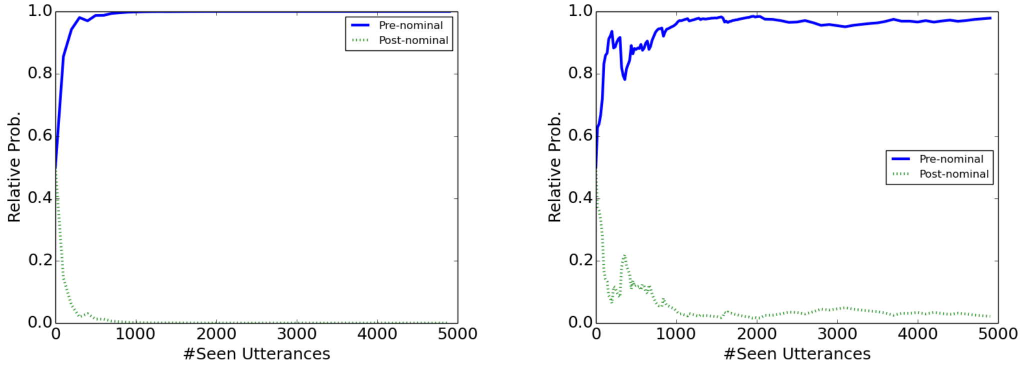
Results with a few other values for the learning rate present similar trends with some changes in the details. Under all settings the correct grammatical categories are favored (and with determiners and prepositions, overwhelmingly so) and the learning of individual words follows similar patterns to those described in the following section. Changes in details include more fluctuating learning curves when the learning rate is more moderate, and somewhat different parameter values upon convergence.

To show that the model also learns the correct word order for other categories, we examined determiners (Figure 4) and prepositions (Figure 5), both in the high-uncertainty settings.[12] (It is uninformative to look at nouns since the logical form in this case uniquely determines the syntactic category, namely N.) Results show that the model quickly converges to the correct options (pre-nominal rather than post-nominal determiners, and prepositions rather than postpositions), even in the presence of high propositional uncertainty. In fact, the learning curves are even steeper than for verbs, despite having no explicit "head-direction" parameter to set.

3.4 Learning Curves of Individual Words

We next examine the learning curves for individual words, again starting with transitive verbs. To do so, we estimate the probability $P(v,\ \mathrm{SVO} \mid m_v)$ (henceforth, "correct production probability" or *CPP*) for a given transitive verb form v and its associated logical form m_v. This probability is a proxy to the production beliefs of the child, i.e., her ability to utter the correct word and syntactic category, given an intended meaning. This is a strict measure, as it requires learning both the correct logical form and the correct syntactic category for the verb. We condition on the logical form instead of the word form in order to avoid penalizing homographs, which are particularly common in our model, which e.g. treats the word "run" in "they run", "run along now", and "I want to run" as having three unrelated syntactic and logical forms which just happen to have the same surface form. This means that even a perfect learner would have $P(m_{\text{third person plural } run} \mid run) < 1$, simply because

[12] We consider only prepositions that introduce an argument rather than an adjunct. Adjunct prepositions have different syntactic and semantic types and are hence excluded by conditioning on the logical form.

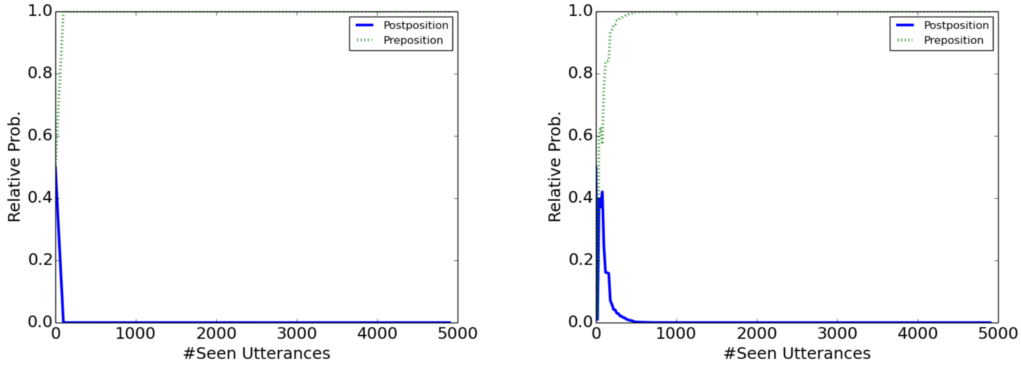


(a) Syntactic disposition for determiners, 4 distractors (b) Syntactic disposition for determiners, 6 distractors

Figure 4. Learning word order for determiners. Plots show the relative probability of generating a pre-nominal vs. a post-nominal determiner in the (a) 4 distractor and (b) 6 distractor setting, as a function of the number of training utterances seen.

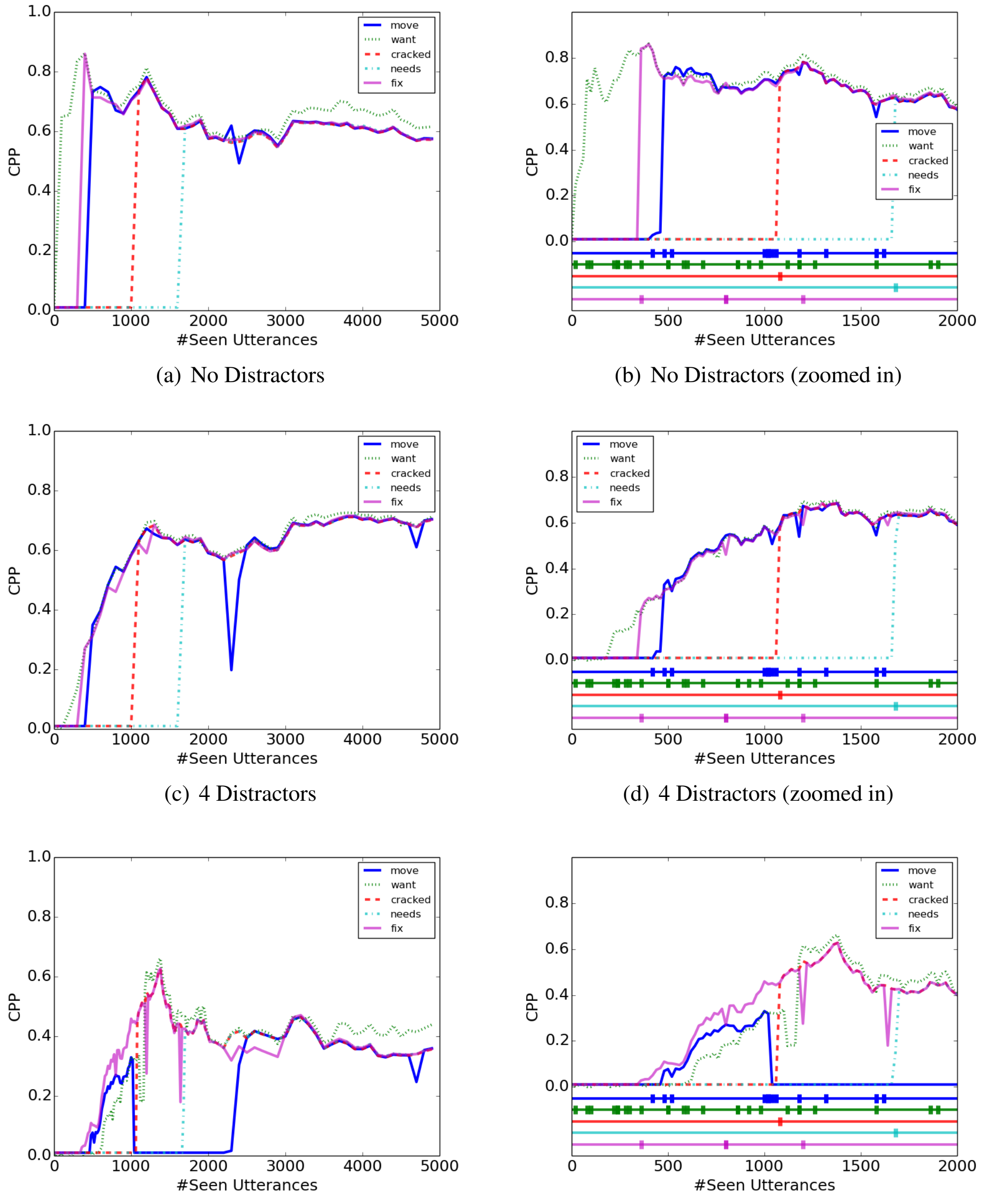
some probability mass must also be reserved for $P(m_{\text{imperative } \textit{run}} \mid \textit{run})$, and this effect would vary in unpredictable ways for different word forms, making it difficult to assess the success of learning. By contrast, in our training data each logical form uniquely determines the correct syntactic type and word form, so that a perfectly successful learner should always approach a CPP of 1.

Figure 6 shows how the CPP varies over the course of training for a few transitive verbs, both frequent (“move”, “want”, “fix”) and less frequent (“cracked”, “needs”). With no propositional uncertainty, all of these verbs are learned rapidly within a short timeframe after they are encountered. In the cases of 4 and 6 distractors, learning of word order is more gradual, revealing how innacurate beliefs about the structure of the language influence the learning of specific verbs. Indeed, verbs that are first observed in an earlier stage of learning (e.g., “want”, “fix”, “move”) no longer exhibit one-shot learning: instead, they are learned only gradually (and even seem to be “un-learned” on



(a) Syntactic disposition for prepositions, 4 distractors (b) Syntactic disposition for prepositions, 6 distractors

Figure 5. Learning word order for prepositions. Plots show the relative probability of generating a preposition vs. a postposition in the (a) 4 distractor and (b) 6 distractor setting, as a function of the number of training utterances seen.



(a) No Distractors

(b) No Distractors (zoomed in)

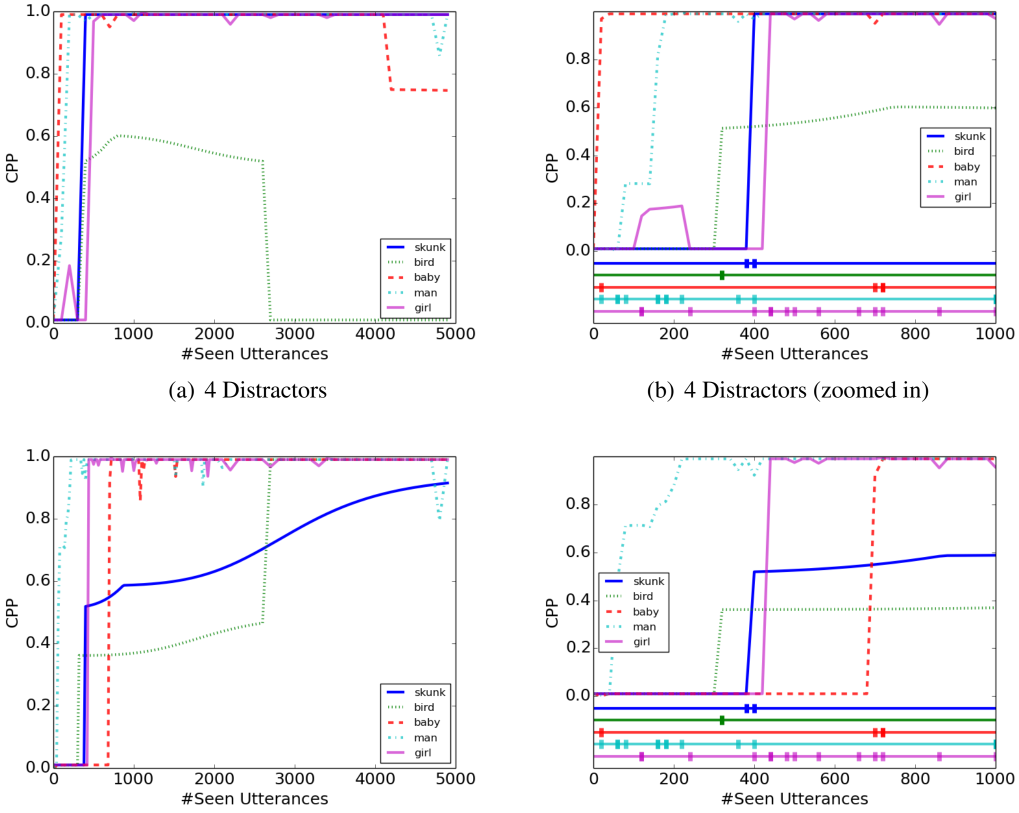
(c) 4 Distractors

(d) 4 Distractors (zoomed in)

(e) 6 Distractors

(f) 6 Distractors (zoomed in)

Figure 6. Learning curves showing the correct production probability for verbs of different frequencies in the (top) 0-distractor, (middle) 4-distractor, and (bottom) 6-distractor settings, as a function of the number of utterances observed. (Results from the 2 distractor setting in this case are uninformative since the model fails to learn SVO word order well: see Figure 2(b) and associated discussion.) Plots on the right zoom in on the first stage of learning, and also show horizontal "timelines" for each word with tickmarks indicating, for each 20-utterance block, whether the word occurred in that block. (The color and top-to-bottom order of the timelines matches the words in the legend.) Large jumps in the plots generally correspond to observations in which the given verb appears, while smaller changes can occur due to fine-tuning of higher-order model parameters (e.g., the overall probability of transitive verbs having SVO syntax). The frequencies for these verbs are: $f(move) = 18$, $f(want) = 70$, $f(cracked) = 1$, $f(needs) = 3$, $f(fix) = 11$.



(a) 4 Distractors

(b) 4 Distractors (zoomed in)

(c) 6 Distractors

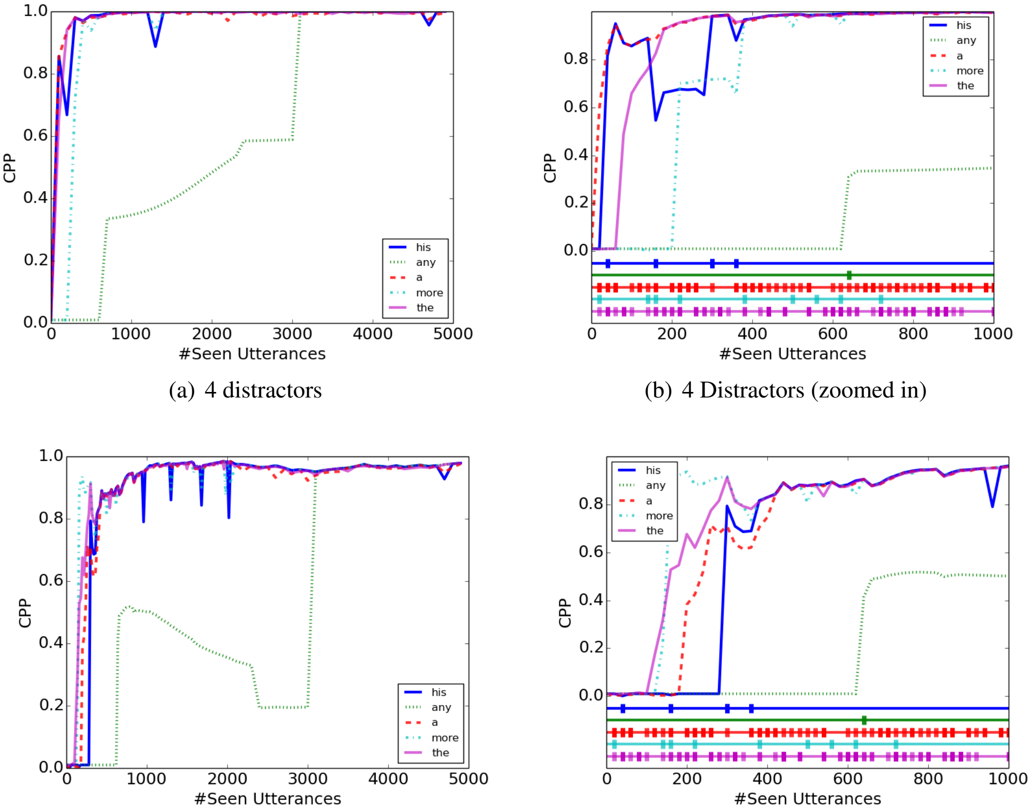
(d) 6 Distractors (zoomed in)

Figure 7. Learning curves showing the CPP for nouns of different frequencies in the (top) 4-distractor and (bottom) 6-distractor settings, as a function of the number of utterances observed. Right-hand plots zoom in on the earlier stage of learning, as in Figure 6. The frequencies are: $f(skunk) = 2, f(bird) = 3, f(baby) = 12, f(man) = 29, f(girl) = 52$.

occasion; though the model is able to recover from the error).

Interestingly, the model's knowledge about these words can actually improve even without seeing them again (see utterances 500-1000 in the 4 and 6 distractor cases) as its implicit higher-order beliefs about grammatical structure (e.g., SVO word order) change. As the model learns more about the language, its acquired knowledge also provides a strong prior guiding the interpretation of utterances containing new verbs. This leads to much more rapid learning of verbs first observed late in learning ("cracked", "needs"). There are two ways in which prior learning can help: first, knowing the meanings of (some) other words in the sentence reduces the uncertainty about the meaning of the verb. Second, as we demonstrate in Section 3.7, even if *no* other words in the sentence are known, simply having high prior probability associated with rules instantiated by SVO categories will help to disambiguate the verb's meaning: syntactic bootstrapping at work. In the naturalistic setting of this simulation, these two sources of knowledge are inextricably linked; Section 3.7 isolates the syntactic effect by simulating a controlled laboratory experiment.

Of course, if our analysis is correct, then these patterns of learning should not be restricted to verbs. Figure 7 presents the CPP for various nouns of different frequencies, showing the high-



(a) 4 distractors

(b) 4 Distractors (zoomed in)

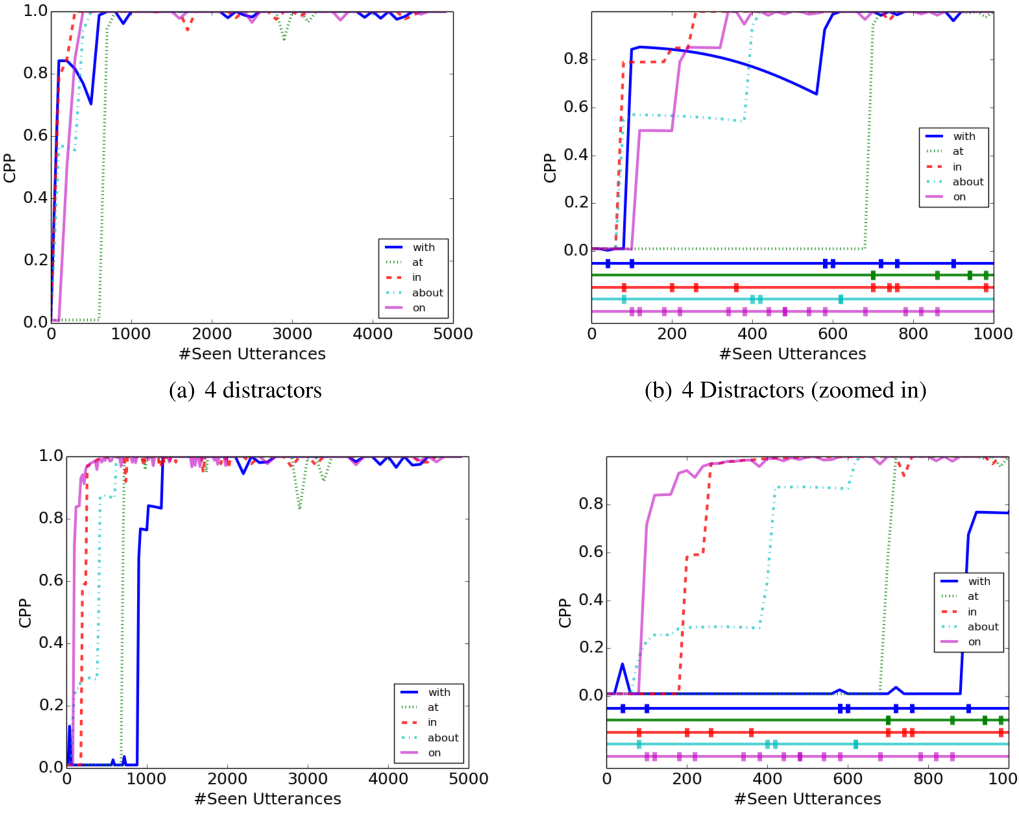
(c) 6 distractors

(d) 6 Distractors (zoomed in)

Figure 8. Learning curves showing the CPP for determiners of different frequencies in the (top) 4-distractor and (bottom) 6-distractor settings, as a function of the number of utterances observed. Right-hand plots zoom in on the earlier stage of learning, as in Figure 6. The frequencies are: $f(his) = 17, f(any) = 9, f(a) = 351, f(more) = 23, f(the) = 321$.

uncertainty scenarios where the differences between early and late learning should be more pronounced. These plots suggest that nouns are easier to learn than verbs (we return to this point in Section 3.6), since most of the nouns are learned rapidly, even early on (note the zoomed in plots only go to 1000 utterances here). Nevertheless, nouns observed in the very earliest part of learning do tend to require more observations to learn than those first observed later: in some cases learning is more gradual (as for “man”); in others the word may be observed a few times with little or no learning, before then being learned all at once (as for “girl”). Figure 8 and Figure 9 show similar effects for some individual determiners and prepositions.

To summarize, results from looking at learning curves for individual words suggest that words first observed early in the learning process are initially learned gradually (or in some cases suddenly but after several occurrences), while those that first appear later on are learned more rapidly, often after only a single occurrence. These findings, which we explore further in the following section, suggest that qualitative changes in the shape of the learning curves can be explained by a single statistical learning mechanism, which improves as it learns (see also Fazly et al. 2010; Regier 2005 and references therein). This single mechanism can also explain the ability to learn from just one



(a) 4 distractors

(b) 4 Distractors (zoomed in)

(c) 6 distractors

(d) 6 Distractors (zoomed in)

Figure 9. Learning curves showing the CPP for prepositions of different frequencies in the (top) 4-distractor and (bottom) 6-distractor settings, as a function of the number of utterances observed. Right-hand plots zoom in on the earlier stage of learning, as in Figure 6. The frequencies are: $f(with) = 53, f(at) = 17, f(in) = 86, f(about) = 7, f(on) = 78$.

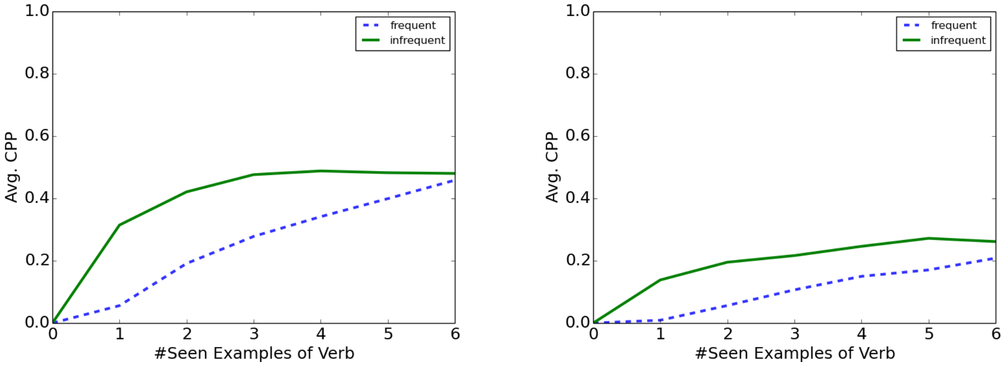
example, a phenomenon we return to in Section 3.7.

3.5 Acceleration of learning and the vocabulary spurt

The learning curves for individual words presented in the previous section suggested that words encountered early in acquisition are learned more slowly than those first encountered at a later point. We now show that this finding is not restricted to the particular words we examined, and we also explore the consequences for vocabulary size as word learning accelerates over time.

For the first analysis, we look at transitive verbs. Notice that, on average, a *high-frequency* verb will appear first earlier in the corpus than a *low-frequency* verb. Thus, if words that first occur late tend to be learned more rapidly, then on average low-frequency verbs will be learned in fewer examples, and high-frequency verbs will require more examples.

To test this hypothesis, we divide verbs into two bins: *high-frequency* (occurs more than 10 times in the corpus) or *low-frequency* (occurs 10 or fewer times). For each bin, we compute the



(a) 4 Distractors

(b) 6 Distractors

Figure 10. The average CPP (y-axis) for frequent and infrequent verbs, after observing *n* examples (x-axis). Because the *n*th example of a low-frequency verb occurs later in training (on average), fewer examples are required to reach a given level of proficiency. We use a frequency threshold of 10, leading to 99 infrequent verbs and 23 frequent verbs. We do not plot 7-9 occurrences because only eight of the infrequent verbs have 7 or more occurrences. Results are shown only for the 4- and 6-distractor settings because in the low-distraction setting, all verbs are learned so quickly that it obscures the effect.

average CPP for all verbs after seeing *n* instances of that verb type.[13] We plot this average CPP as a function of *n* in Figure 10, which indeed shows the predicted trend.

Our second analysis, which focuses on nouns, is more directly comparable to observations from real children. Specifically, after an initial period in which the vocabulary of the child is restricted to no more than a few words, a sudden increase in the production vocabulary size is observed, usually around the age of 18–24 months (Reznick & Goldfield, 1992). The trends we have already observed in individual words, when taken together over the whole vocabulary, should yield exactly this kind of accelerated learning.

Figure 11 confirms this prediction (see Fazly et al. 2010 for a related finding). We plot an estimate of the size of the noun production vocabulary as a function of the number of observed utterances. The noun vocabulary size is defined as the number of nouns whose CPP exceeds a threshold of 0.8. We present results for the 6 distractors setting in which the trends can be seen most clearly, as the initial learning stage is longer. Results show that after an initial period (up to about 200 utterances) in which the vocabulary size increases slowly, the slope of the curve changes and nouns are learned at a faster rate. This could be partly due to syntactic bootstrapping from knowledge about determiner and preposition ordering, which is acquired very early (as shown in Figures 4 and 5). It is likely also due to the fact that as more nouns are learned, any sentence containing those nouns has fewer other unknown words, so overall uncertainty is reduced. In any case, the results suggest that no special mechanism or change in processing is needed to explain the vocabulary spurt, it simply arises from the nature of the statistical learner.

[13] Many verbs are ambiguous between multiple logical forms. We therefore define a verb type according to its m_v rather than its written form.

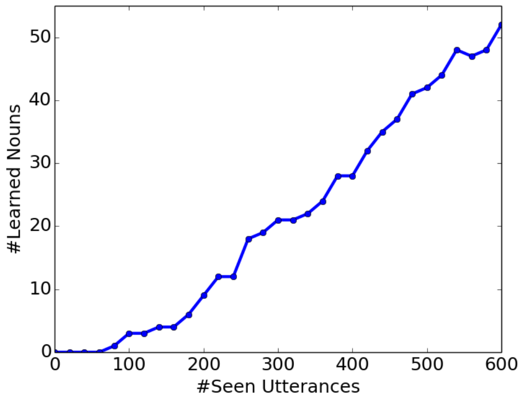


Figure 11. The size of the production vocabulary of nouns as a function of the number of utterances seen, for the 6 distractors setting. The production vocabulary is defined as all nouns with a CPP of at least 0.8.

3.6 Comparing noun and verb learning

It is well-known that nouns predominate over verbs in the early lexicons of English-learning children, although there is considerable individual and cross-linguistic variation, and some verbs or predicative categories are among the earliest vocabulary items acquired for some children (Gentner, 1982; Tomasello, 1992:210). Figure 12 shows that our learner shows a similar trend, accumulating nouns faster than transitive verbs, but learning some verbs right from the start.[14]

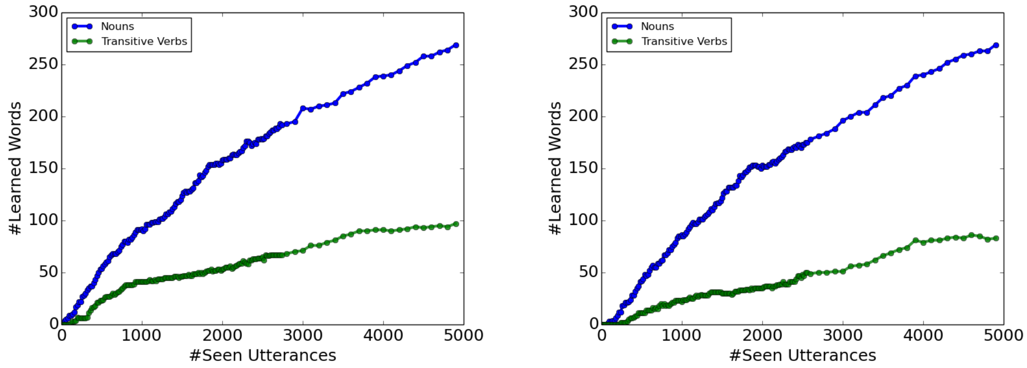
This observed bias towards nouns cannot be ascribed to a difference in the conceptual complexity of nouns and verbs, or a difference in their level of concreteness (see Snedeker & Gleitman, 2004 for a review of proposals based on these ideas), as our model does not distinguish nouns and verbs on any such dimensions.

Moreover, the larger noun vocabulary in our model doesn't simply reflect the input frequencies of nouns and verbs: Figure 13 presents the ratio of verb types to noun types in the input, as well as in the production vocabulary for the different distractor settings. In the more realistic 2-, 4- and 6-distractor settings, the verb:noun ratio starts off *lower* than that in the input, and only later converges to the input ratio. That is, early on (especially in the first few hundreds of sentences), verbs are under-represented relative to their input rate. In the no-distractors setting, the learned rate is similar to the input rate.[15] This under-representation of verbs in the early phase of learning is consistent with claims of noun bias in the literature, i.e., that the the ratio of verbs to nouns is lower in child productions than in the child directed speech (see discussion in Gentner & Boroditsky, 2001).

Comparing the plots in Figure 13 with the learning curves for word order shown earlier suggests an explanation for the noun bias in our model, namely that: (1) early on, nouns are over-

[14] There are many different ways to count word types, e.g., deciding whether to lump or split a verb by subcategorization, tense, agreement, mood and other grammatical systems may result in different counts. Here we focus on transitive verbs, counting types of all inflections, while counting only singular nouns against the noun's vocabulary. Thus, any difference between the noun and transitive verb vocabulary sizes is a conservative estimate. These methodological issues play a lesser role when comparing the ratios of the input and learned vocabularies (see below).

[15] The plot appears to show that verbs are over-represented for the first 80 utterances, but as the vocabulary is very small at this point (no more than ten words of each category), the difference between the learned and input ratios is likely not statistically significant.



(a) 4 Distractors (b) 6 Distractors

Figure 12. A comparison of the noun and transitive verb vocabulary sizes in the (a) 4-distractor and (b) 6-distractor settings. The curves present the sizes of the production vocabularies of nouns and transitive verbs (defined as words with CPP at least 0.8) as a function of the number of observed utterances. Results indicate that nouns account for a larger part of the vocabulary than transitive verbs. The relative vocabulary sizes for the 0- and 2-distractor settings are nearly identical to what is shown, but more words are learned in the 0-distractor setting.

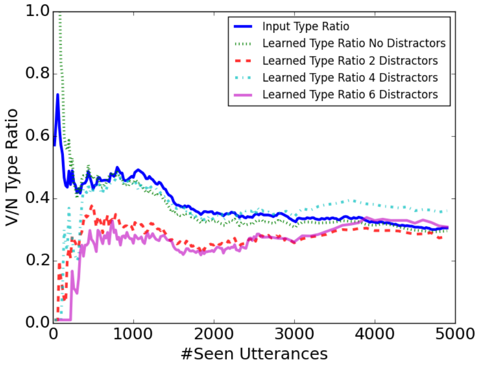


Figure 13. A comparison of the ratio of transitive verbs and nouns in the child-directed input, relative to this ratio in the child's production vocabulary (defined as words with CPP at least 0.8). Curves are presented for all four distractor settings. Results indicate that the ratios converge to similar values in all settings, but that in the 2, 4 and 6 distractor settings verbs are initially under-represented relative to their presence in the input.

represented in the vocabulary because the syntactic constructions of noun phrases (involving determiners and prepositions) are learned more quickly than those of transitive clauses (chiefly, the category of transitive verbs), allowing syntactic bootstrapping to start earlier for nouns and increase the rate of noun learning; (2) as learning progresses, and once the structure of transitive clauses is learned too, the rate of verb learning "catches up" to that of nouns, and the larger noun vocabulary can be directly ascribed to the differing frequencies of verbs and nouns in child-directed speech. Indeed, the point of acquisition of the correct SVO category in the different settings (Figure 2) largely matches the later phase of convergence between the input and learned rate. The no-distractor setting thus does not display an early under-representation of verbs simply because the syntax of transitives is learned very early on.

This explanation is indeed consistent with the proposal of Snedeker and Gleitman (2004) that verbs lag behind nouns because learning verbs requires greater *linguistic* (rather than *conceptual*) development on the part of the child. In particular, they argue that verb learning requires more knowledge about the surrounding sentential context (including its syntax) so the learner gets additional cues beyond just the non-linguistic situational context.

Beyond this hypothesis, our model also suggests that the use of syntactic cues might kick in earlier for nouns than for verbs. Whether this holds true in infants likely requires further behavioral study, since most behavioral work on syntactic bootstrapping has focused on verbs (see Fisher, Gertner, Scott, & Yuan, 2010 for a review). Nevertheless, some studies do suggest that infants are able to use syntactic and morphological cues earlier in noun processing than verb processing (see Cauvet et al., 2014 for a review). It is therefore plausible that syntactic information relevant to nouns is learned and used earlier, and that the resulting bootstrapping effect could further explain the early noun bias in children's vocabulary.

3.7 Syntactic bootstrapping for one-shot learning of nonce words

So far, all our analyses have focused on the "natural" timecourse of learning, studying what the simulated child might know at a particular moment in time as a result of their particular and idiosyncratic history of exposure to language. Much of the literature on child language learning, however, focuses on what children of different ages are able to learn based on controlled exposure to a small number of examples in a laboratory setting. In this section, we follow suit, by first training our model on the first *n* examples from the corpus (to simulate the background knowledge a child of a particular age might bring to the lab), and then exposing it to a small number of hand-crafted training examples and studying what it learns from these.

In particular, we examine the model's ability to deal with cases like Fisher et al.'s (1994) *chase*/*flee* example, where the same physical situation seems to support more than one logical form. Fisher et al. (1994) ask how it is that children faced with (made up) examples like the following avoid the error of making an OVS lexical entry for "flee" with the meaning *chase*:

(7) "Kitties flee doggies"

It is important that words like "flee", where the instigator of the event is realized as an object rather than a subject, are relatively rare. (For example, there are eight occurrences of any form of "flee" in the entire English CHILDES corpus—over 17M tokens—of which none is transitive.) Therefore, the child will have encountered plenty of normal transitive verbs (where the instigator is the subject) before encountering examples like (7).

In terms of our model, this means that, by the time (7) is encountered, the learned prior probability of the instantiated rules for combining transitive SVO verbs with their object and subject will be substantially greater than the priors for OVS verbs. This will lead to a much higher probability for the correct meaning and SVO syntax for “flee” than for an OVS lexical item with the same meaning as “chase”. Of course, the model also entertains other possibilities, such as that “flee” means *cats*, that “kitties” means *flee*, etc. However, assuming that “kitties” and “doggies” have been encountered relatively often, so that “flee” is the only unfamiliar word in the sentence, then the model will assign a very low probability to these additional spurious hypotheses.

In fact, we will see that after training, the model can determine the correct meaning for a transitive verb even in the face of significantly greater ambiguity, when the meanings of other words in the sentence are *not* known, because the learned prior over the OVS syntactic category, which is not attested elsewhere in the data, becomes exceedingly low.

Gleitman and colleagues have described the process by which the child resolves this type of contextual ambiguity as “syntactic bootstrapping”, meaning that lexical acquisition is guided by the child’s knowledge of the language-specific grammar, as opposed to the semantics (Gleitman, 1990; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). However, in present terms such an influence on learning simply emerges from the statistical model required for semantic bootstrapping. We will return to this point in the General Discussion.

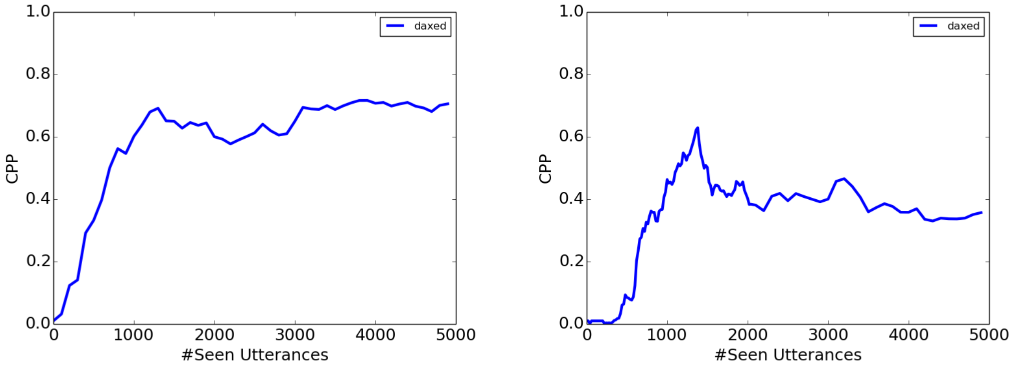
3.7.1 Nonce verbs. Our first simulation is inspired by the experiments of Gertner, Fisher, and Eisengart (2006), where children at the age of 21 and 25 months were presented with two pictures, one in which a girl is performing some action over a boy, and one in which it is the other way around. The pictures were paired with a description, either “the boy gorped the girl” or “the girl gorped the boy”. Results showed a disposition of the children to select the interpretation (that is, to look at the picture) in which the actor coincided with the subject rather than the object of the sentence.

Here, we examine the model’s ability to correctly interpret the nonce verb “dax” in the sentence “the man daxed the baby”, when presented with two possible meaning representations: one in which the man is performing an action over the baby, and one in which the baby is performing an action over the man. In order to correctly infer the meaning of “daxed” (analogous to Fisher et al’s “flee” and Gertner et al.’s “gorped”) the model has to rely on its learned prior to favor SVO over OVS interpretations, as the two interpretations are otherwise symmetrical. Results show the model quickly learns to favor the correct SVO interpretation for “daxed”.

Specifically, we first train the model on the first n examples of the CHILDES corpus, in intervals of 100 examples (4 and 6 distractors settings), and then present the model with one more sentence – literally, the string “the man daxed the baby” – which is paired with both the SVO- and OVS-supporting logical forms. (Intuitively, the SVO-supporting logical form involving the man daxing the baby is ‘correct’, while the OVS-supporting logical form involving the baby daxing the man is a ‘distractor’, but the model is simply presented both as two a priori equally likely possibilities.) We then query the model’s state to compute several predictions.

First, we ask which of the two logical forms the model assigns a higher posterior probability, and find that from $n = 200$ examples and onwards the correct SVO interpretation is always preferred. Next, we use the strict CPP statistic to assess the model’s success at learning the nonce verb “daxed” : $\lambda x \lambda y.daxed(y,x)$ based on this single training example (Figure 14) and find that it is quite successful despite the syntactic complexity of the surrounding sentence.

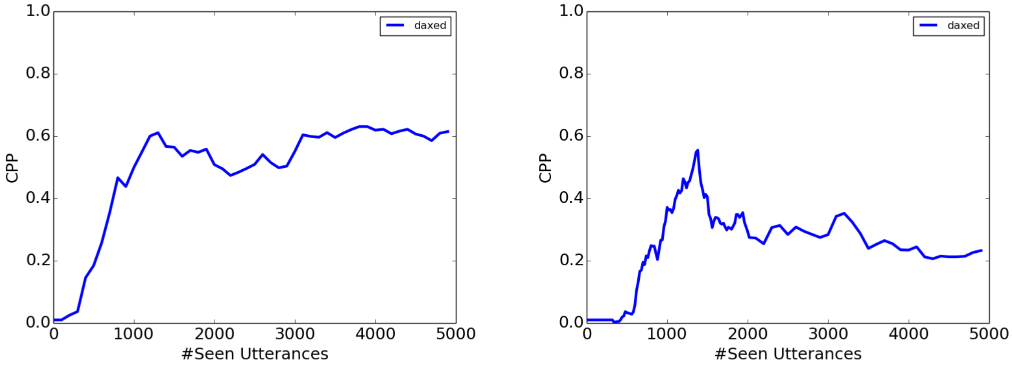
3.7.2 Multiple novel words. The above simulation examined the model’s ability to perform syntactic bootstrapping to learn an unknown verb in the presence of uncertainty between an



(a) “The man daxed the baby”, 4 distractors

(b) “The man daxed the baby”, 6 distractors

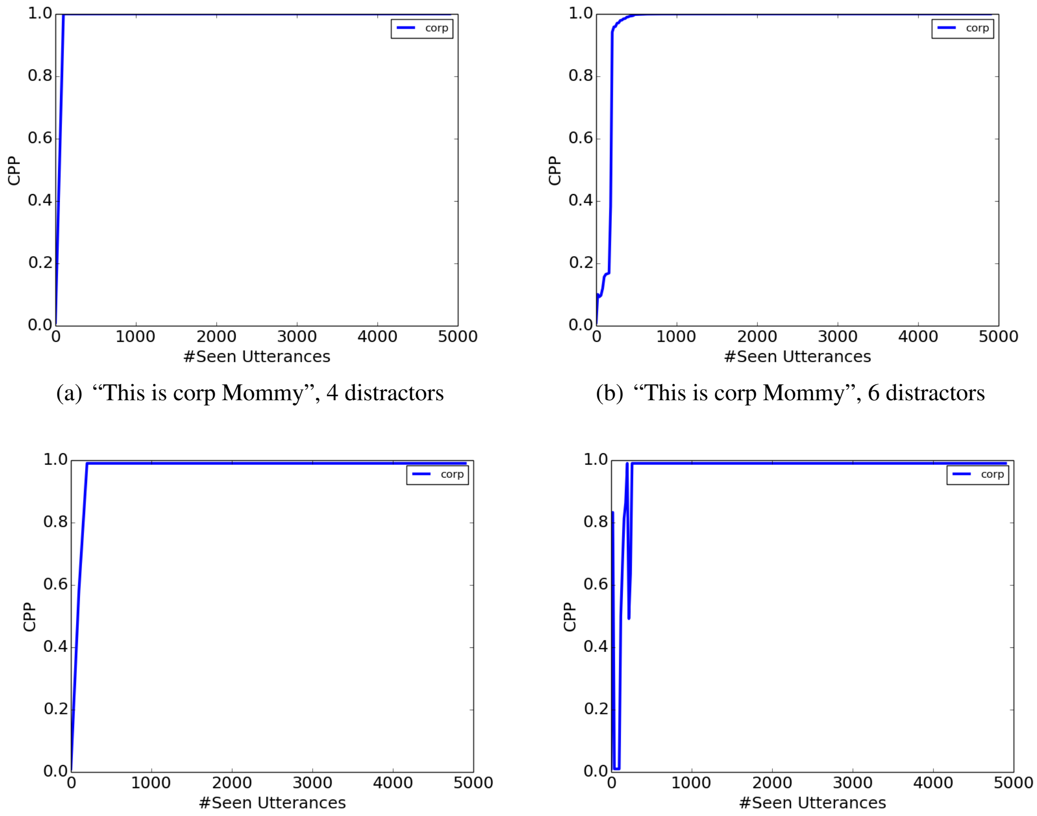
Figure 14. Syntactic bootstrapping: The model’s CPP for the unknown word “daxed” (y-axis) after training on *n* utterances from the corpus (x-axis) followed by a single presentation of the sentence “The man daxed the baby” with SVO/OVS ambiguous semantics. During training only, either four (Figure 14(a)) or six (Figure 14(b)) distractors are included along with each correct logical form.



(a) “Jacky daxed Jacob”, 4 distractors

(b) “Jacky daxed Jacob”, 6 distractors

Figure 15. Syntactic bootstrapping: The model’s CPP for the word “daxed” (y-axis) after training on *n* utterances from the corpus (x-axis) plus one exposure to “Jacky daxed Jacob”, where all three words are unknown. During training only, either four (Figure 15(a)) or six (Figure 15(b)) distractors are included along with each correct logical form. In both cases, the correct interpretation of the test sentence is the most likely one upon convergence.



(a) “This is corp Mommy”, 4 distractors

(b) “This is corp Mommy”, 6 distractors

(c) “This is the corp”, 4 distractors

(d) “This is the corp”, 6 distractors

Figure 16. The correct production probability (y-axis) for the unknown word “corp” in the sentences “This is corp Mommy”, where it should be interpreted as a preposition (16(a) and 16(b)), and “This is the corp” (16(c) and 16(d)), where it should be interpreted as a noun, as a function of the number of training examples seen (x-axis). The model is trained in the 4 and 6 distractors settings.

SVO and an OVS interpretation. We next examine the model’s performance in a simulated experiment where the utterance is ambiguous between all possible word orders. We find that also in this setting, the model is able to determine the correct SVO order, due to its high (learned) prior over the rules instantiated by the SVO category as opposed to the alternatives.

Concretely, in this simulation the model is trained on n examples from the corpus, and then receives the single utterance “Jacky daxed Jacob”, where “Jacky”, “Jacob” and “daxed” are all unknown words, paired with the logical form $daxed(jacky, jacob)$. This corresponds to a situation where the child is exposed to a scene in which some hitherto unnamed person performs an unusual action on another unnamed person. A priori, there are many possible interpretations that are consistent with this utterance; for example, in a VSO language we would expect a child to interpret “Jacky” as a verb meaning $daxed(x, y)$, and “daxed” as a noun meaning *jacky*, and if we exposed our model to this example alone without any other training data, then it would indeed assign this interpretation just as high a probability as the intended SVO interpretation. What this simulated experiment tests, therefore, is the model’s ability to make higher-order syntactic generalizations from the English-based training corpus, and apply them to this novel utterance.

First, we ask which derivation tree for “Jacky daxed Jacob” : $daxed(daxer, daxee)$ is assigned the highest probability, and find that the correct SVO derivation is preferred in all cases where $n \geq 500$ for the 4 distractor case and $n \geq 800$ for the 6 distractor case. We further examine the model’s ability to assign “daxed” the correct SVO interpretation by plotting the CPP of the word “daxed” after training on *n* training examples, followed by the dax example. Figure 15(a) and (b) present this probability for the 4 and 6 distractor settings respectively. These results demonstrate the model’s ability to correctly interpret “daxed” (and consequently the sentence) in this more ambiguous setting after training on a single example of the verb, proving the model’s ability to learn and apply syntactic generalizations even in the presence of propositional uncertainty.

3.7.3 Nonce nouns and prepositions. Finally, we examine the cases of nouns and prepositions, confirming the model’s ability to perform syntactic bootstrapping in these categories as well. We take inspiration from Fisher et al. (2006), who demonstrated 2-year old’s disposition to interpret “corp” in “this is a corp” as a noun (and therefore assign it a noun-compatible meaning), while interpreting “acorp” in “this is acorp the baby” as a preposition (with a preposition-like meaning). In our case, we feed the model with either of the two utterances: “this is the corp” and “this is corp Mommy”, paired with two possible logical forms: one where “corp” is compatible with the noun category and another where “corp” is compatible with the preposition category.[16] The model assigns “corp” a noun interpretation in the first example, and a preposition interpretation in the second. Concretely, as before the model is trained on *n* examples in the 4 and 6 distractors settings, and then given a single instance of either of the “corp” examples, paired with the two possible logical forms. The steep learning curves of this simulation are given in Figure 16, indicating that once again the model is able to use its knowledge of syntax to correctly disambiguate between the meanings, just as the children in the experiment did.

4 General Discussion

We have presented an incremental model of language acquisition that learns a probabilistic CCG grammar from utterances paired with one or more potential meanings. The learning model assumes no knowledge specific to the target language, but does assume that the learner has access to a universal functional mapping from syntactic to semantic types (E. Klein & Sag, 1985), as well as a Bayesian model favoring grammars with heavy reuse of existing rules and lexical types. It is one of only a few computational models that learn syntax and semantics concurrently, and the only one we know of that both (a) learns from a naturalistic corpus and (b) is evaluated on a wide range of phenomena from the acquisition literature. Moreover, the model is fully generative, meaning that it can be used for both parsing (understanding) and generation, and that probability distributions over subsets of the variables (e.g., individual lexical items) are well-defined—a property we used in many of our analyses. Like some very recent proposals (Beekhuizen, 2015; Jones, 2015) but unlike earlier work (Buttery, 2006; Gibson & Wexler, 1994; Sakas & Fodor, 2001; Siskind, 1992; Villavicencio, 2002; Yang, 2002), we also evaluate our learner by parsing sentences onto their meanings, rather than just examining the learner’s ability to set a small number of predefined syntactic parameters.

Together, these properties of our model and evaluation allow us to investigate the relationship between word learning and syntactic acquisition in a more realistic setting than previous work.

[16] The slight modifications of Fisher et al.’s examples were done in order to arrive at a pair of utterances of the same length, in which the non-nonce words are learned very early on.

Results from our simulations suggest that the kind of probabilistic semantic bootstrapping model proposed here is sufficient to explain a range of phenomena from the acquisition literature: acceleration of word learning, including one-shot learning of many items later in acquisition; the predominance of nouns over verbs in early acquisition; and the ability to disambiguate word meanings using syntax. All of these behaviors emerge from the basic assumptions of the model; no additional stipulations or mechanisms are required. Finally, our work addresses a criticism levelled against some statistical learners—that their learning curves are too gradual (Thornton & Tesan, 2007). By demonstrating sudden learning of word order and of word meanings, our model shows that statistical learners can account for sudden changes in children's grammars.

In the remainder of the discussion, we expand on a few of these points, highlight some connections between our model and earlier theories of acquisition, and discuss how our model might generalize to more realistic meaning representations.

4.1 Syntactic bootstrapping revisited

It has long been noted that the onset of syntactically productive language is accompanied by an explosion of advances in qualitatively different "operational" cognitive abilities (Vygotsky, 1934/1986). These findings suggest that language has a feedback effect that facilitates access to difficult concepts, a view that is further supported by the early work of Oléron (1953) and Furth (1961), who demonstrated that deaf children who were linguistically deprived (by being denied access to sign) showed specific cognitive deficits in non-perceptually evident concepts.

This view is also consistent with Gleitman's (1990) proposal that the availability of syntax enables the child to "syntactically bootstrap" lexical entries for verbs (such as "promise") that may not be situationally evident. However, we have argued that such syntactic bootstrapping is only possible when the child has access to a model of the relation between language-specific syntax and the universal conceptual representation. Our simulations of learning novel transitive verbs, inspired by the *chase/flee* ambiguity and related behavioral studies, support this argument and demonstrate that syntactic bootstrapping need not be postulated as a distinct mechanism, but is emergent from a semantic bootstrapping approach such as the one proposed here.

In that sense, the so-called "syntactic" bootstrapping effect might better be termed "structural", since it relies crucially on (and arises directly from) the homomorphism between syntactic and semantic types. Although the child might eventually learn the meanings of novel predicates in any case, structural bootstrapping serves to highlight the meaning that is most plausible in light of the child's previously acquired grammatical knowledge. Our model operationalizes that previously acquired knowledge as a probabilistic grammar, where the probabilities learned from previous examples act as the prior distribution when faced with a new example. When learning from a naturalistic corpus, these priors become sufficiently peaked that the learner is able to identify the correct verb meanings in examples with ambiguous semantic roles, like "the man daxed the baby". More generally, the learner's accumulating syntactic knowledge (along with knowing more word meanings) causes an acceleration of learning, which we demonstrated both for verbs and nouns.

4.2 Relationship to Non-probabilistic and Parameter-Setting Approaches

Like other probabilistic approaches to language acquisition, our proposal avoids many of the stipulations required by non-probabilistic approaches. In particular, there is no need for a Subset Principle (Berwick, 1985) or the ordered presentation of unambiguous parametric triggers, both of which appear to present serious problems for the language learner (Angluin 1980; Becker 2005; J. D. Fodor and Sakas 2005). Nor does our approach contradict widely-held assumptions concerning the lack of negative evidence in the child's input. The child can progress from the universal superset grammar to the language-specific target grammar entirely on the basis of positive evidence. This positive evidence raises the probability of correct hypotheses at the expense of incorrect ones, including those introduced by error and noise. The only evidence that the child needs is a reasonable proportion of utterances that are sufficiently short for them to deal with.

Nevertheless, our proposal does share some commonalities with earlier non-probabilistic approaches. For example, it resembles the proposal of J. D. Fodor (1998) as developed in Sakas and Fodor (2001) and P. Niyogi (2006) in that it treats the acquisition of grammar as arising from parsing with a universal "supergrammar", which includes all possible syntactic structures allowed in any language that are compatible with the data (J. A. Fodor, Bever, and Garrett 1974:475). However, our learner uses a different algorithm to converge on the target grammar. Rather than learning rules in an all or none fashion on the basis of parametrically unambiguous sentences, it uses *all* processable utterances to adjust the probabilities of all elements of the grammar for which there is positive evidence. In this respect, it more closely resembles the proposal of Yang (2002). However it differs from both in eschewing the view that grammar learning is parameter setting, and from the latter in using a Bayesian learning model rather than Mathematical Learning Theory (Atkinson, Bower, & Crothers, 1965).

We eschew parameter setting because if the parameters are implicit in the rules or categories themselves, which can be learned directly, neither the child nor the theory need be concerned with parameters other than as a prior distribution over the seen alternatives. For the child, all-or-none parameter-setting would be counterproductive, making it hard to learn the many languages that have exceptional lexical items or inconsistent settings of parameters across lexical types, as in the English auxiliary verb system and Italian subject-postposing intransitives. Of course, languages do tend toward consistent (if violable) values of parameters like headedness across categories for related semantic types (such as verbs and prepositions). Such consistency will make learning easier under the present theory, by raising the model's probabilities for recurring rules and categories, leading to higher prior probabilities for those rules and categories when they are re-encountered in future. It is less clear that representing this consistency explicitly, rather than implicitly in the model and the evolving prior, will help the individual child learning its first language.

4.3 Cognitive plausibility and interpretation of probabilities

Our model is based on similar mathematical principles to other Bayesian models of language acquisition (e.g., Feldman et al., 2013; M. Frank et al., 2009; Goldwater et al., 2009; Griffiths & Tenenbaum, 2005, 2006; Perfors, Tenenbaum, & Wonnacott, 2010; Xu & Tenenbaum, 2007), and shares with them the attractive properties noted above: the probabilistic framework permits the learner to accumulate evidence over time and eventually make strong inferences about the language-specific grammar, without the need for negative evidence and in a way that is robust to noise. However, unlike many Bayesian models, our learner uses an *incremental* learning algorithm rather than adopting an ideal observer approach. We adopted this approach in the interests of cognitive plausibility and in order to easily examine the progress of learning over time. As a result our learner has no guarantee of optimality, but is aligned with other recent work investigating cognitively plausible approximate learning strategies for Bayesian learners (Pearl, Goldwater, &

Steyvers, 2010; Sanborn, in press; Sanborn, Griffiths, & Navarro, 2010; Shi, Griffiths, Feldman, & Sanborn, 2010).

Nevertheless it is worth noting that, despite the incremental algorithm, some other aspects of our learner are less cognitively plausible. In particular, to process each input example the learner generates all parses consistent with that example. This can require significant memory and computation, especially for longer sentences, and indeed is not incremental at the level of words, as human sentence processing is generally accepted to be. However, it might be possible to develop further approximations to reduce the computational requirements, for example by computing parameter updates based on a high probability subset of the possible parses. Another approach might be particle filtering, which has been used by some other recent models to model memory limitations in incremental sentence processing and other tasks (Levy, Reali, & Griffiths, 2009; Sanborn et al., 2010; Shi et al., 2010).

Regardless of the learning algorithm, the use of probabilities raises the question of how exactly a notion like "the adult grammar" should be understood in this model, where the grammar represents a probability distribution over all hypotheses that the child learner has ever entertained, albeit one that has converged with most of the probability mass concentrated in correct lexical entries and rules of derivation, as opposed to spurious determiners like "doggies" and rules for combining them with nouns like "more".

Such a grammar/model could in principle be pruned to eliminate such obvious dead wood. However, the grammar would remain probabilistic in nature. This conclusion might seem to conflict with traditional theories of competence, but such probabilistic grammars are now widespread in models of gradience in language processing and gradient grammaticality (Bresnan & Nikitina, 2003; Hale, 2001; Jurafsky, 1996; Levy, 2008; Sorace & Keller, 2005). Some have argued that gradience in competence is distinct from that in performance (Crocker & Keller, 2005). Our model's probabilities closely reflect processing (i.e., frequency of use). However, even if interpreted as a model of competence, it can in principle model both gradient and hard judgements, the latter either as extremely low-probability events, or as structures that are entirely absent from the probabilistic grammar. Our model would then instantiate a theory of learning for such gradient grammars, analogous to various learning models proposed for stochastic Optimality Theoretic grammars (e.g., Boersma & Hayes, 2001; Goldwater & Johnson, 2003; Hayes & Wilson, 2008).

4.4 Towards More Realistic Meaning Representation

The meaning representations for child-directed utterance considered in this work have been necessarily simplified in content, because the CHILDES annotations that we have to work with are limited to narrow, propositional literal meaning. However, the content that the child learns from surely also includes aspects of the social interaction that the child is engaged in (Tomasello & Farrar, 1986, *passim)*. It is the latter that provides the primary motivation for the child to learn language, particularly in the earliest stages. Such interpersonal content is lacking in the CHILDES annotation.

We know of no fully convincing or widely accepted formal semantic representation for these interpersonal aspects of discourse meaning, much less any such representation that we believe human annotators could apply consistently to labelled corpora of child directed utterance (Calhoun et al., 2010; Cook & Bildhauer, 2011).

However, we do believe that the knowledge representation identified by Tomasello can be expressed as compositional semantic formulas, not fundamentally different from the ones we use here to demonstrate the workings of our learner. Indeed, Steedman (2014) proposed such a compositional formulation for one major component of interactional learning, namely information structure.

The term "information structure" refers to the packaging of the content of an utterance along a number of dimensions that have been referred to in the literature—not always consistently—as "topic", "comment", "contrast" and the like, together with markers of speaker and hearer orientation of these meaning components (Bolinger, 1965; Halliday, 1967; Ladd, 1996; Steedman, 2014), often expressed by prosodical markers such as pitch contour and metrical alignment (Calhoun, 2010, 2012). Intonational prosody is used exuberantly in speech by and to children, and we know that infants are sensitive to the interpersonal significance of intonation from a very early age (Fernald, 1993; Fernald et al., 1989).

It is therefore likely that when a child hears her mother pointing at newly observed dogs and saying "More doggies!", the child's representation of the utterance includes the information that a new entity or property is introduced to the common ground, and that she considers the possibility that this information is conveyed by the speaker's intonation contour. Using the intonational notation of Pierrehumbert and Hirschberg (1990) (H* and LL% are a high accent and low boundary tone respectively) and sub-categorizing the syntactic categories to correspond to information-structural distinctions, the derivation might look as follows ($+,\rho$ marks a positive polarity rheme or comment, and ϕ marks a complete phonological phrase):

(8)
$$\begin{array}{cccl}
\text{M}\textsc{ore} & \textsc{doggies} & ! & \\
\text{H}* & \text{H}* & \text{LL\%} & \\
\hline
NP_{+,\rho}/N_{+,\rho} & N_{+,\rho} & X_{\phi}\backslash_{\star}X_{\pi,\eta} & \\
\cline{1-2}
\multicolumn{2}{c}{NP_{+,\rho}} & & > \\
\cline{1-3}
\multicolumn{3}{c}{NP_{\phi}} & <
\end{array}$$

The semantics of this sentence can be glossed as: "Mum makes the property afforded by more dogs common ground". The details of the semantics of information structure go beyond the immediate concerns of this paper: the example is merely intended to suggest that interpersonal aspects of meaning can be treated as formally as more standard content.

Intonation structure does not always align with traditional syntactic structure even in adult dialog, and the two diverge even more in child-directed and child-originated speech (Fisher & Tokura, 1996; Gerken, 1996; Gerken, Jusczyk, & Mandel, 1994). However, in CCG, intonation structure is united with a freer notion of derivational structure. Consider the child in a similar situation to the above, who hears example (1) from Section 2.1 "You like the doggies!" with the intonation contour discussed by Fisher herself and by Steedman (1996a):[17]

[17] This derivation crucially involves *type-raising* of the subject *You*. Type raising simply exchanges the roles of function and argument, making the subject apply to the predicate to give the same result. Although, in the interests of simplification, we have glossed over the point in the preceding discussion of the learner, *all* NPs are always type-raised in this sense by a lexical process analogous to morphological *case*, or what the linguists call "structural" case, so the object is also type-raised. Thus, this example involves the CCG equivalent of *nominative* and *accusative* case. The responsibility for deciding which structural case is relevant lies with the parsing model, rather than the grammar.

(9)

$$\begin{array}{ccccl}
\text{You} & \text{LIKE} & & \text{the doggies!} & \\
 & \text{H}* & \text{L} & \text{LL\%} & \\
\overline{S/(S\backslash NP)} & \overline{(S\backslash NP)/NP} & \overline{X_{\phi}\backslash_{\star}X_{\pi,\eta}} & \overline{S_{\phi}\backslash(S_{\phi}/NP_{\phi})} & \\
\cline{1-2}
\multicolumn{2}{c}{S/NP} & {>}\mathbf{B} & & \\
\cline{1-3}
\multicolumn{3}{c}{S_{\phi}/NP_{\phi}} & {<} & \\
\cline{1-4}
\multicolumn{4}{c}{S_{\phi}} & {<}
\end{array}$$

"Mum supposes the properties the dogs afford to be common ground, Mum makes it common ground it's what I like."

Fisher points out that the L intermediate phrase boundary that she observed after the verb makes the intonation structure inconsistent with standard assumptions about English NP-VP surface constituency. However, CCG embodies a more permissive notion of constituency, allowing for constituents such as "you like", consistent with the intonation structure above. As the derivation requires the use of the forward composition rule, indicated as $>\mathbf{B}$, the intonation contour provides positive evidence for applying the composition rule to the first two categories, supporting syntactic learning from prosodic cues. Far from being inconsistent with the syntax of English, the intonation structure is aligned to the notion of constituent structure that the child will later find essential to the analysis of relative clause structures like the following:

(10) [The [doggie [that **[you like]**$_{S/NP}$]$_{N\backslash N}$]$_{N}$]NP

Thus, the child can acquire the relative construction without needing to call on a distinct and otherwise unmotivated mechanism of movement. This strong homomorphism between interpretation and syntactic constituency motivates the use of CCG as the model of grammar in a theory of language acquisition.

It is also worth noting that even the content that *is* expressed in our input meaning representations is inevitably more shallow and English-specific than we would like, even though it abstracts away from English-specific word order. For example, comparison with other Germanic languages suggests that the logical form that the child brings to (9) may be something more like $give(pleasure(you), dogs)$, so that the lexical entry for "like" of type $(e,(e,t))$ is the following, exhibiting the same "quirky" relation between (structural) nominative case and an underlying dative role that Icelandic exhibits morphologically for the corresponding verb:

(11) like := $(S\backslash NP)/NP : \lambda x\lambda y.give(pleasure(y), x)$

However, dealing with these further issues is beyond the scope of the present investigation, because of the small amount of data we have to work with, and the restricted nature of the annotation available to us.

5 Conclusion

This paper has presented, to our knowledge, the first model of child language acquisition that jointly learns both word meanings and syntax and is evaluated on naturalistic child-directed sentences paired with structured representations of their meaning. We have demonstrated that the model reproduces several important characteristics of child language acquisition, including rapid learning of word order, an acceleration in vocabulary learning, one-shot learning of novel words, and syntactic bootstrapping effects.

Taken together, these results support our basic claim that syntax is learned by semantic bootstrapping from contextually available semantic interpretations using a statistical learning mechanism that operates over a universally available set of grammatical possibilities. This learning mechanism gradually shifts probability mass towards frequent candidates, allowing the child to rapidly acquire an (approximately) correct grammar even in the face of competing ambiguous alternative meanings. Altogether, the model offers a unified account of several behavioral phenomena, demonstrating in particular that both syntactic bootstrapping effects and changes in the pace of learning emerge naturally from the kind of statistical semantic bootstrapping model presented here—no additional mechanisms are needed.

Acknowledgements

The work was supported in part by ERC Advanced Fellowship 249520 GRAMPLUS, EU IST Cognitive Systems IP EC-FP7-270273 Xperience, ARC Discovery grants DP 110102506 and 160102156, and James S McDonnell Foundation Scholar Award 220020374. We thank Julia Hockenmaier, Mark Johnson, and Luke Zettlemoyer for guidance in the early stages of this research, and Inbal Arnon, Jennifer Culbertson, and Ida Szubert for their feedback on a draft of this article.

References

Abend, O., Reichart, R., & Rappoport, A. (2010). Improved unsupervised POS induction through prototype discovery. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1298–1307).

Alishahi, A., & Chrupała, G. (2012). Concurrent acquisition of word meaning and lexical categories. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (p. 643-654). Jeju Island.

Alishahi, A., Fazly, A., & Stevenson, S. (2008). Fast mapping in word learning: What probabilities tell us. In *Proceedings of the twelfth conference on computational natural language learning* (pp. 57–64).

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*, 789-834.

Alishahi, A., & Stevenson, S. (2010). A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, *25*(1), 50–93.

Allen, J., & Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist networks. *The emergence of language*, 115–151.

Ambati, B. R., Deoskar, T., & Steedman, M. (2016). Hindi CCGbank: A CCG treebank from the Hindi dependency treebank. *Language Resources and Evaluation*, *50*, submitted.

Ambridge, B., Pine, J., & Lieven, E. (2014). Child language acquisition: Why universal grammar doesn't help. *Language*, *90*, e53–e90.

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, *45*, 117-135.

Artzi, Y., Das, D., & Petrov, S. (2014). Learning compact lexicons for CCG semantic parsing. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (p. 1273-1283).

Atkinson, R., Bower, G., & Crothers, E. (1965). *Introduction to mathematical learning theory*. Wiley.

Auli, M., & Lopez, A. (2011). A comparison of loopy belief propagation and dual decomposition for integrated CCG supertagging and parsing. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (p. 470-480). Portland, OR: ACL.

Baldridge, J. (2002). *Lexically specified derivational control in Combinatory Categorial Grammar* (Unpublished doctoral dissertation). University of Edinburgh.

Barak, L., Fazly, A., & Stevenson, S. (2013). Modeling the emergence of an exemplar verb in construction learning. In *Proceedings of the 35th annual conference of the cognitive science society.* Berlin.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* (Unpublished doctoral dissertation). University of London.

Becker, M. (2005). Raising, control, and the subset principle. In *Proceedings of the 24th west coast conference on formal linguistics* (p. 52-60). Somerville, MA: Cascadilla Proceedings Project.

Beekhuizen, B. (2015). *Constructions emerging: a usage-based model of the acquisition of grammar* (Unpublished doctoral dissertation). Leiden University.

Beekhuizen, B., Bod, R., Fazly, A., Stevenson, S., & Verhagen, A. (2014). A usage-based model of early grammatical development. In *Proceedings of the ACL workshop on cognitive modeling and computational linguistics.* Baltimore, MD.

Berwick, R. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, *32*(1), 45–86.

Bolinger, D. (1965). *Forms of English*. Cambridge, MA: Harvard University Press.

Bowerman, M. (1973). Structural relationships in children's utterances: Syntactic or semantic? In *Cognitive development and the acquisition of language.* Academic Press.

Braine, M. (1992). What sort of innate structure is needed to "bootstrap" into syntax? *Cognition*, *45*, 77-100.

Bresnan, J. (Ed.). (1982). *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Bresnan, J., & Nikitina, T. (2003). *On the gradience of the dative alternation.* Unpublished manuscript. Stanford University.

Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Brown, R., & Bellugi, U. (1964). Three processes in the child's acquisition of syntax. In E. Lenneberg (Ed.), *New directions in the study of language* (p. 131-161). Cambridge, MA: MIT Press.

Buttery, P. (2006). *Computational models for first language acquisition* (Unpublished doctoral dissertation). University of Cambridge.

Calhoun, S. (2010). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, *86*, 1-42.

Calhoun, S. (2012). The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics*, *40*, 329-349.

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., & Beaver, D. (2010). The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics, and prosody of dialog. *Language Resources and Evaluation*, *44*, 387-419.

Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., & Christophe, A. (2014). Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Language Learning and Development*, *10*(1), 1–18.

Çakıcı, R. (2005). Automatic induction of a CCG grammar for Turkish. In *Proceedings of the acl student workshop* (p. 73-78). ACL.

Chang, N. C.-L. (2008). *Constructing grammar: A computational model of the emergence of early constructions*. ProQuest.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th national conference of the american association for artificial intelligence* (p. 598-603).

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2010). Two decades of unsupervised POS tagging—how far have we come? In *Proceedings of the conference on empirical methods in natural language processing* (p. 575-584). ACL.

Chrupała, G., Kádár, Á., & Alishahi, A. (2015). Learning language through pictures. In *Proceedings of the 53nd annual meeting of the association for computational linguistics* (pp. 112–118).

Clark, E. (1973). What's in a word? on the child's acquisition of semantics in his first language. In T. Moore (Ed.), *Cognitive development and the acquisition of language* (p. 65-110). Academic Press.

Clark, S., & Curran, J. R. (2004). Parsing the WSJ using CCG and log-linear models. In *Proceedings of the*

42nd annual meeting of the association for computational linguistics (p. 104-111). Barcelona, Spain: ACL.

Cohn, T., Blunsom, P., & Goldwater, S. (2010). Inducing tree-substitution grammars. *The Journal of Machine Learning Research*, *9999*, 3053-3096.

Collins, M. (1997). Three generative lexicalized models for statistical parsing. In *Proceedings of the 35th annual meeting of the association for computational linguistics* (p. 16-23). Madrid: ACL.

Connor, M., Fisher, C., & Roth, D. (2012). Starting from scratch in semantic role labeling: Early indirect supervision. In *Cognitive aspects of computational language acquisition* (p. 257-296). Springer.

Cook, P., & Bildhauer, F. (2011). Annotating information structure: The case of topic. In *Beyond semantics: Corpus based investigations of pragmatic and discourse phenomena* (p. 45-56). Ruhr Universität, Bochum: Bochumer Linguistische Arbeitsberichte.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 522-543.

Crocker, M. W., & Keller, F. (2005). Probabilistic grammars as models of gradience in language processing. In *Gradience in grammar: Generative perspectives.* Oxford, UK: Oxford University Press.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Culbertson, J., Smolensky, P., & Wilson, C. (2013). Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, *5*, 392-424.

Dominey, P. F., & Boucher, J.-D. (2005). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1), 31–61.

Elman, J., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017-1063.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Development*, *64*, 657-667.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to infants. *Journal of Child Language*, *16*, 477-501.

Fisher, C., Gertner, Y., Scott, R., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 143-149.

Fisher, C., Hall, G., Rakowitz, S., & Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, *92*, 333-375.

Fisher, C., Klingler, S. L., & Song, H.-j. (2006). What does syntax say about space? 2-year-olds use sentence structure to learn new prepositions. *Cognition*, *101*(1), B19-B29.

Fisher, C., & Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (p. 343-363). Erlbaum.

Fodor, J. A., Bever, T., & Garrett, M. (1974). *The psychology of language.* New York: McGraw-Hill.

Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry*, *29*, 1-36.

Fodor, J. D., & Sakas, W. (2005). The subset principle in syntax: Costs of compliance. *Journal of Linguistics*, *41*, 513-569.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578-585.

Frank, S., Feldman, N., & Goldwater, S. (2014). Weak semantic context helps phonetic learning in a model of infant language acquisition. In *Proceedings of the 52nd annual meeting of the association of computational linguistics.*

Furth, H. (1961). The influence of language on the development of concept formation in deaf children. *Journal of Abnormal and Social Psychology*, *63*, 386-389.

Gazdar, G., Klein, E., Pullum, G. K., & Sag, I. (1985). *Generalized phrase structure grammar*. Oxford: Blackwell.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language development* (Vol. 2, p. 301-334). Hillsdale, NJ: Erlbaum.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. In M. Bowerman & S. Levinson (Eds.), (p. 215-256). Cambridge: Cambridge University Press.

Gerken, L. (1996). Prosodic structure in young children's language production. *Langauge*, *72*, 683-712.

Gerken, L., Jusczyk, P., & Mandel, D. (1994). When prosody fails to cue syntactic structure. *Cognition*, *51*, 237-265.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*(8), 684–691.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*, 355-407.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 1-55.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*, 23-64.

Göksun, T., Küntay, A. C., & Naigles, L. R. (2008). Turkish children use morphosyntactic bootstrapping in interpreting verb meaning. *Journal of Child Language*, *35*, 291-323.

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21-54.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the workshop on variation within optimality theory* (pp. 113–122). Stockholm University.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*(2), 183–206.

Griffiths, T., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334-384.

Griffiths, T., & Tenenbaum, J. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767 -773.

Grimshaw, J. (1981). Form, function and the language acquisition device. In L. Baker & J. McCarthy (Eds.), *The logical problem of language acquisition* (p. 165-182). Cambridge, MA: MIT Press.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd meeting of the north american chapter of the association for computational linguistics* (p. 159-166). Pittsburgh, PA.

Halliday, M. (1967). *Intonation and grammar in British English*. The Hague: Mouton.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379–440.

Hockenmaier, J. (2003). Parsing with generative models of predicate-argument structure. In *Proceedings of the 41st meeting of the association for computational linguistics, sapporo* (p. 359-366). San Francisco: Morgan-Kaufmann.

Hockenmaier, J., & Steedman, M. (2002). Generative models for statistical parsing with Combinatory Categorial Grammar. In *Proceedings of the 40th meeting of the association for computational linguistics* (p. 335-342). Philadelphia.

Hoffman, M., Blei, D., & Bach, F. (2010). Online learning for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, *23*, 856-864.

Hyams, N. (1986). *Language acquisition and the theory of parameters*. Dordrecht: Reidel.

Johnson, M., & Goldwater, S. (2009). Improving nonparametric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of human language technologies: The 2009 annual conference of the north American chapter of the association for computational linguistics.*

Jones, B. K. (2015). *Learning words and syntactic cues in highly ambiguous contexts* (Unpublished doctoral dissertation). University of Edinburgh.

Joshi, A., & Schabes, Y. (1997). Tree-Adjoining Grammars. In G. Rozenberg & A. Salomaa (Eds.), *Handbook of formal languages* (Vol. 3, p. 69-124). Berlin: Springer.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Klein, D., & Manning, C. D. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd annual meeting of the association for computational linguistics* (p. 479-486). Barcelona: ACL.

Klein, D., & Manning, C. D. (2005). Natural language grammar induction with a generative constituent-context model. *Pattern Recognition*, *38*, 1407-1419.

Klein, E., & Sag, I. A. (1985). Type-driven translation. *Linguistics and Philosophy*, *8*, 163-201.

Krishnamurthy, J., & Mitchell, T. (2014). Joint syntactic and semantic parsing with combinatory categorial grammar. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (p. 1188-1198). Baltimore, MD.

Kwiatkowski, T. (2012). *Probabilistic grammar induction from sentences and structured meanings* (Unpublished doctoral dissertation). University of Edinburgh.

Kwiatkowski, T., Goldwater, S., Zettlemoyer, L., & Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th conference of the european chapter of the ACL (EACL 2012)* (p. 234-244). Avignon: ACL.

Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., & Steedman, M. (2010). Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the conference on empirical methods in natural language processing* (p. 1223-1233). Cambridge, MA: ACL.

Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., & Steedman, M. (2011). Lexical generalization in CCG grammar induction for semantic parsing. In *Proceedings of the conference on empirical methods in natural language processing* (p. 1512-1523). Edinburgh: ACL.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press. (2nd edition revised 2008)

Landau, B., & Gleitman, L. (1985). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126-1177.

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In *Proceedings of the 22nd conference on neural information processing systems (nips).*

Lewis, M., & Steedman, M. (2014). A* CCG parsing with a supertag-factored model. In *Proceedings of the conference on empirical methods in natural language processing* (p. 990-1000). Doha, Qatar: ACL.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

Matuszek, C., Fitzgerald, N., Zettlemoyer, L., Bo, L., & Fox, D. (2012). A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th international conference on machine learning (icml).*

Maurits, L., Perfors, A., & Navarro, D. (2009). Joint acquisition of word order and word reference. In *Proceedings of the 31st annual conference of the cognitive science society* (p. 1728-1733).

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological review*, *119*(4), 831.

Mellish, C. (1989). Some chart-based techniques for parsing ill-formed input. In *Proceedings of the 27th annual meeting of the association for computational linguistics* (p. 102-109).

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech 2010, 11th annual conference of the international speech communication association* (pp. 1045–1048).

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (p. 746-751). Atlanta: ACL.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91-117.

Morris, W. C., Cottrell, G. W., & Elman, J. (2000). A connectionist simulation of the empirical acquisition of grammatical relations. In *Hybrid neural systems* (pp. 175–193). Springer.

Niyogi, P. (2006). *Computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Niyogi, P., & Berwick, R. (1996). A language learning model for finite parameter spaces. *Cognition*, *61*, 161-193.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In *Proceedings of the 24th annual conference of the cognitive science society* (Vol. 36, p. 58-63).

Oléron, P. (1953). Conceptual thinking of the deaf. *American Annals of the Deaf*, *98*, 304-310.

Pearl, L., Goldwater, S., & Steyvers, M. (2010). How ideal are we? incorporating human limitations into Bayesian models of word segmentation. In *Proceedings of the 34th annual boston university conference on child language development.* Somerville, MA: Cascadilla Press.

Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*, 306-338.

Perfors, A., Tenenbaum, J., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*, 607-642.

Phillips, L., & Pearl, L. (2014). Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. In *Proceedings of the computational and cognitive models of language acquisition and language processing workshop at EACL.*

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication* (p. 271-312). Cambridge, MA: MIT Press.

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*, 217-283.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, *4*(3-4), 293–312.

Pollard, C., & Sag, I. (1994). *Head driven phrase structure grammar*. Stanford, CA: CSLI Publications.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425-469.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive science*, *29*(6), 819–865.

Reznick, J. S., & Goldfield, B. A. (1992). Rapid change in lexical development in comprehension and production. *Developmental psychology*, *28*(3), 406.

Ross, J. R. (1967). *Constraints on variables in syntax* (Unpublished doctoral dissertation). MIT. (Published as Ross 1986)

Ross, J. R. (1986). *Infinite syntax!* Norton, NJ: Ablex.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of childes transcripts. *Journal of Child Language*, *37*, 705-729.

Sakas, W., & Fodor, J. D. (2001). The structural triggers learner. In S. Bertolo (Ed.), *Language acquisition and learnability* (p. 172-233). Cambridge: Cambridge University Press.

Sanborn, A. (in press). Types of approximation for probabilistic cognition: sampling and variational. *Brain and Cognition*.

Sanborn, A., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144-1167.

Sato, M.-A. (2001). Online model selection based on the variational Bayes. *Neural Computation*, *13*(7), 1649-1681.

Schlesinger, I. (1971). Production of utterances and language acquisition. In D. Slobin (Ed.), *The ontogenesis of grammar* (p. 63-101). New York: Academic Press.

Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. (2010). Exemplar models as a mechanism for performing

Bayesian inference. *Psychonomic bulletin & review*, *17*(4), 443–464.

Siskind, J. (1992). *Naive physics, event perception, lexical semantics, and language acquisition* (Unpublished doctoral dissertation). MIT.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39-91.

Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. In G. Hall & S. Waxman (Eds.), *Weaving a lexicon* (p. 257-294). MIT Press.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*, 1497-1524.

Steedman, M. (1996a). The role of prosody and semantics in the acquisition of syntax. In J. Morgan & K. Demuth (Eds.), *Signal to syntax* (p. 331-342). Hillsdale, NJ: Erlbaum.

Steedman, M. (1996b). *Surface structure and interpretation*. Cambridge, MA: MIT Press.

Steedman, M. (2000). *The syntactic process*. Cambridge, MA: MIT Press.

Steedman, M. (2012). *Taking scope: The natural semantics of quantifiers*. Cambridge, MA: MIT Press.

Steedman, M. (2014). The surface-compositional semantics of English intonation. *Language*, *90*, 2-57.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Proceedings of interspeech* (pp. 194–197).

Thomforde, E., & Steedman, M. (2011). Semi-supervised CCG lexicon extension. In *Proceedings of the conference on empirical methods in natural language processing* (p. 1246-1256). ACL.

Thompson, C., & Mooney, R. (2003). Acquiring word-meaning mappings for natural language interfaces. *Journal of Artificial Intelligence Research*, *18*, 1-44.

Thornton, R., & Tesan, G. (2007). Categorical acquisition: Parameter setting in Universal Grammar. *Biolinguistics*, *1*, 49-98.

Tomasello, M. (1992). *First verbs: a case study in early grammatical development*. Cambridge: Cambridge University Press.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M., & Farrar, M. (1986). Joint attention and early language. *Child development*, 1454-1463.

Trueswell, J., & Gleitman, L. (2007). Learning to parse and its implications for language acquisition. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (p. 635-656). Oxford: Oxford University Press.

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141-188.

Ural, A. E., Yuret, D., Ketrez, F. N., Koçbaş, D., & Küntay, A. C. (2009). Morphological cues vs. number of nominals in learning verb types in Turkish: The syntactic bootstrapping mechanism revisited. *Language and Cognitive Processes*, *24*, 1393-1405.

Villavicencio, A. (2002). *The acquisition of a unification-based generalised categorial grammar* (Unpublished doctoral dissertation). University of Cambridge.

Vygotsky, L. (1934/1986). *Thought and language*. MIT Press: Cambridge, MA. (Trans. A. Kozulin))

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, C. (2006). *The infinite gift*. New York: Scribner.

Yu, C. (2006). Learning syntax–semantics mappings to bootstrap word learning. In *Proceedings of the 28th annual conference of the cognitive science society* (p. 924-929).

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149-2165.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, *8*(11), e79659.

Yu, H., & Siskind, J. (2013). Grounded language learning from video described with sentences. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (p. 53-63). Sofia.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: the baby's view is better. *Developmental science*, *16*(6), 959–966.

Zettlemoyer, L., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with Probabilistic Categorial Grammars. In *Proceedings of the 21st conference on uncertainty in AI (UAI)* (p. 658-666). Edinburgh: AAAI.

Zettlemoyer, L., & Collins, M. (2007). Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning* (p. 678-687). Prague: ACL.

Appendix A
The Probabilistic Model

Recall the probabilistic model, defining a family of distributions over (s,m,t) triplets consisting of an utterance, meaning representations for the leaves, and a derivation tree. The model is an instance of a Probabilistic Context-free Grammar (PCFG), and decomposes over the derivation steps (or expansions). Specifically, let $w_1,...,w_n$ be the words in the sentence s, $m_1,...,m_n$ be their meaning representations and $c_1,...,c_n$ be their syntactic categories (the syntactic categories at the leaves of t). The probability of an (s,m,t) triplet is:

$$(10) \qquad P(s,m,t) = \prod_{a\to\gamma\in t} P(\gamma \mid a) \cdot \prod_{i=1}^{n} P(m_i \mid c_i) \cdot \prod_{i=1}^{n} P(w_i \mid m_i)$$

Defining the model consists of defining the distribution of the different expansions: the distribution $P(\gamma \mid a)$ over each syntactic expansion γ (consisting of one or more nonterminal syntactic categories) given its parent nonterminal category a; the distribution $P(m_i \mid c_i)$ over each leaf meaning representation given its (terminal) syntactic category; and the distribution $P(w_i \mid m_i)$ over each word given its meaning representation. We refer to all these cases collectively as *expansions* of the parent category (a, c_i or m_i). The distributions of the expansions are defined using Dirichlet Processes (DPs). A DP can serve as the prior over a discrete distribution with infinite support, and (as noted in Section 2.3) yields the following posterior distribution over expansions:

$$(11) \qquad P(\gamma \mid a) = \frac{n_{a\to\gamma} + \alpha_{syn} H_a}{n_a + \alpha_{syn}}$$

where $n_{a\to\gamma}$ is the number of times the expansion $a \to \gamma$ has been used previously and n_a is the number of times any expansion headed by a has been used (in fact, since we never actually observe any parses, the algorithm we describe in Appendix C defines $n_{a\to\gamma}$ and n_a as the *expectations* of these numbers). The model has two hyperparameters: the base distribution H_a and concentration parameter α_a, discussed below. We use here the notation for syntactic expansions, but the same equation applies to the expansion of a category into meaning representation $c \to m$ or meaning representation into wordform $m \to w$.

A Dirichlet Process is therefore defined by the concentration parameter and base distribution (which are stipulated in this work), and the sufficient statistics $n_{a\to b}$ for every a,b, which are estimated from the training data. As only a finite number of (a,b) pairs are non-zero at any given point of the inference procedure, we store the pseudo-counts for these pairs explicitly. The following subsections define the base distributions and concentration parameters used in each part of the model.

Syntactic Expansions

First we define the base distribution of syntactic expansions. We begin by defining an auxiliary distribution over the space of syntactic categories, i.e., the set of all categories that can be defined using brackets and slashes over the set of atomic categories. For a category c, we define $n_{slashes}$ as the number of slash symbols in c (e.g., $n_{slashes}$ of an atomic category is 0 and $n_{slashes}(S\backslash NP/NP)$ is 2). Then:

$$Pr_{cats}(c) \propto 0.2^{n_{slashes}(c)+1} \tag{12}$$

While this does not define a proper distribution over all syntactic categories of any length (as $\Sigma_c Pr_{cats}(c)$ is unbounded), it can be normalized if we assume the vocabulary and formula length are bounded, resulting in a well-defined distribution.

The possible expansions from a syntactic category X depend on its identity. We first define $n_{op}(\mathsf{X})$ as the number of possible CCG inverted operators that can apply to X. Any category X can expand using right and left inverted application, resulting in $(\mathsf{X/Y}, \mathsf{Y})$ and $(\mathsf{Y}, \mathsf{X\backslash Y})$. If X is complex (e.g., $\mathsf{X} = \mathsf{X_1/X_2}$), it can also be expanded through inverted composition (e.g., to $\mathsf{X_1/Y}$ and $\mathsf{Y/X_2}$). In general, if X is complex and takes its first k arguments to its right (left), it can be expanded through right (left) inverted composition of orders $1, ..., k$. n_{op} can be defined accordingly. For each of the allowable operators, the expansion is fully determined by the identity of the introduced category (Y in the examples above). In addition, a category may always be expanded into a lexical item by generating the symbol X_{LEX}. H_a for a syntactic category a is then defined as:

$$H_a(b) \propto \begin{cases} 0.5 \cdot C(b) & b = a_{LEX} \\ \frac{Pr_{cats}(Y)}{2 \cdot n_{op}} & \text{b is a syntactic expansion, introducing } \mathsf{Y} \end{cases} \tag{13}$$

$C(b)$ equals 1 if $n_{slashes}(b) < 2$, equals $\frac{4}{3}$ if b's innermost slashes are in the same direction (both arguments to the right or both to the left, as in $\mathsf{X/Y/Z}$), and equals $\frac{2}{3}$ otherwise (innermost slashes in different directions, as in $\mathsf{X/Y\backslash Z}$). This term corrects the base distribution to be uniform over the six possible constituent orders for a transitive verb after the exclusion of the two cases where the verb attaches first to the first semantic argument (see Section 2.2).

Meaning Representations

The distribution of the expansion of leaf categories into meaning representations $P(m_i \mid c_i)$ is done in two steps. First, c_i generates a shell-meaning representation, which does not contain any constants but instead has (typed) placeholders. Another generation step is then used to expand the shells into fully-fledged meaning representations. For instance, to generate the meaning representation $m_i = \lambda x.cookie(x)$, the corresponding leaf c_i first generates $m_i^{sh} = \lambda x.P(x)$, and then generates m_i given m_i^{sh}.

The reason for performing this two-step derivation is to allow the model to learn a disposition towards sub-classes of logical forms. For instance, assume the learner is exposed to a Verb-Subject-Object language, and correctly learns that the category $\mathsf{S/NP/NP}$ is more likely than the other transitive categories. Now assume the learner faces a transitive clause, in which none of the words are known $\mathsf{w}_1\mathsf{w}_2\mathsf{w}_3 : p(arg_1, arg_2)$. The model will correctly assign high probability to analyses where w_1 corresponds to p, but without the additional $c_i \rightarrow m^{sh}$ will be unable to determine whether w_2 corresponds to arg_1 and w_3 to arg_2 or the other way around. Recall that the VSO and VOS derivations differ only in the meaning representation assigned to w_1 (Section 2.2). Specifically, VSO involves the expansion $\lambda x \lambda y.p(x,y) \rightarrow w_1$ and VOS the expansion $\lambda y \lambda x.p(x,y) \rightarrow w_1$. Both of these derivations have not been previously observed and are symmetrical. A single derivation step would therefore assign them equal probabilities, regardless of whether other learned transitive verbs have VSO or VOS lexical entries.

Using the intermediate shell generation step addresses this problem. If the observed data better attests for VSO than VOS verbs, the derivation $S/NP/NP \rightarrow \lambda x \lambda y.PLACE\_HOLDER(x,y)$ would gain a higher probability mass than $S/NP/NP \rightarrow \lambda y \lambda x.PLACE\_HOLDER(x,y)$. Consequently, the VSO interpretation for w_1 would be preferred over the VOS.

Both derivation steps $c_i \rightarrow m_i^{sh}$ and $m_i^{sh} \rightarrow m_i$ are modelled as DPs[18]. The base distribution in both cases is similarly defined[19]:

$$H_{sem}(m) \propto e^{\#sym + 2\#vars} \tag{14}$$

where $\#sym$ is the number of constants, logical operators and quantifiers appearing in m, and $\#vars$ is the number of variable types (not instances) appearing in m. Again this is an improper distribution, but can be normalized if we assume the vocabulary and formula length are bounded.

The inclusion of shell logical forms in the derivation process serves an additional purpose when inferring meaning representations for sentences s that contain unknown words. While accurate inference is impossible in this case, as the model has no information what logical constants correspond to the unknown words, it is able to infer the most likely logical form, where for unknown words w_i, m^{sh} is inferred instead of m_i. This may result in a meaning representation for the sentence that contains place-holders instead of some of the logical constants, but is otherwise accurate.

Wordforms

The base distribution of the generation of words from logical types $m_i \rightarrow w_i$ is defined as follows:

$$H_{word}(w) = 0.002^{|w|} \tag{15}$$

where $|w|$ is the number of characters in w.

[18] We use the same DP for c_i's that only differ in the slash direction. For instance, $c_i = S/NP$ and $c_i = S\backslash NP$ share a distribution of $P(\cdot \mid c_i)$. This was done to prevent an artefact where unknown meaning representations generated from an infrequent category receive disproportionally large probability mass relative to their probability for being generated from a frequent category. This modification introduces an additional simplifying independence assumption (namely that the distribution of meaning representations does not depend on the directionality of the corresponding leaf syntactic category), which we intend to review in future work.

[19] Much of the space over which the base distribution is defined cannot be attested for due to type constraints. For instance, where the leaf category is of type $c_i = \mathsf{NP}$, the resulting meaning representation must be of type e. Nevertheless, we use this base distribution uniformly for its simplicity.

Syntactic Category	Semantic Type	Example Phrase
$\mathsf{S_{dcl}}$	$\langle ev, t\rangle$	I took it $\vdash$ $\mathsf{S_{dcl}} : \lambda e.took(i,it,e)$
$\mathsf{S_t}$	t	I'm angry $\vdash$ $\mathsf{S_t} : angry(i)$
$\mathsf{S_{wh}}$	$\langle e, \langle ev, t\rangle\rangle$	Who took it? $\vdash$ $\mathsf{S_{wh}} : \lambda x \lambda e.took(x,it,e)$
$\mathsf{S_q}$	$\langle ev, t\rangle$	Did you take it? $\vdash$ $\mathsf{S_q} : \lambda e.Q(take(you,it,e))$
N	$\langle e, t\rangle$	cookie $\vdash$ $\mathsf{N} : \lambda x.cookie(x)$
NP	e	John $\vdash$ $\mathsf{NP} : john$
PP	$\langle ev, t\rangle$	on John $\vdash$ $\mathsf{PP} : \lambda e.on(john,e)$

Figure B1. Atomic Syntactic Categories and their corresponding semantic types.

Concentration parameters

The concentration parameters we use are $\alpha_{syn} = 1$ for the syntactic expansions, $\alpha_{sh} = 1000$ for the expansions of syntactic categories to shells, $\alpha_m = 500$ for the expansions of shells to meaning representations, and $\alpha_w = 1$ for the expansion of meaning representations to words.

Appendix B
The Space of Possible Derivations

The set of allowed parses $\mathbf{t}$ is defined by the function $\mathscr{T}$, mapping pairs of strings and meaning representations (s,m) into a set of possible derivations. Here we review the *splitting procedure* of Kwiatkowski et al. (2010) that is used to generate CCG lexical items and describe how it is used by $\mathscr{T}$ to create a packed chart representation of all parses $\mathbf{t}$ that are consistent with (s,m).

The splitting procedure takes as input a CCG category $\mathsf{X} : h$, such as $\mathsf{NP} : a(x, cookie(x))$, and returns a set of *category splits*. Each category split is a pair of CCG categories $(\mathsf{C_l} : m_l, \mathsf{C_r} : m_r)$ (ordered by their linear order) that can be recombined to give $\mathsf{X} : h$ using one of the CCG combinators in Section 2.2. The CCG category splitting procedure has two parts: *logical splitting* of the category semantics h; and *syntactic splitting* of the syntactic category X. Each logical split of h is a pair of lambda expressions (f,g) in the following set:

(16) $$\{(f,g) \mid h = f(g) \ \vee \ h = \lambda x.f(g(x))\},$$

which means that f and g can be recombined using either *function application* or *function composition* to give the original lambda expression h. A number of further linguistically motivated constraints on possible splits apply, such as an "Across the Board" constraint that says that abstraction over a term A must apply to all instances of A, and an "A over A" constraint that says you only consider the topmost term of type A for abstraction, and not any terms of the same type A that it may contain (cf. Ross 1967). These are similar to the constraints described in Kwiatkowski (2012, :116–118). An example split of the lambda expression $h = a(x, cookie(x))$ is the pair

(17) $$(\lambda y.a(x,y(x)), \lambda x.cookie(x)),$$

where $\lambda y.a(x,y(x))$ applied to $\lambda x.cookie(x)$ returns the original expression $a(x, cookie(x))$.

If f and g can be recombined to give h with function application, and X is the syntactic category of h, then X is split by a reversal of the CCG application combinators in Section 2.2, yielding the following set of possible categories for f and g:

(18) $$\{(X/Y : f \ \ Y : g), (Y : g \ \ : X\backslash Y : f)\}$$

where $X : h$ and $h = f(g)$. Similarly, if f and g can be recombined to give h with function composition and if $X = X_1/X_2$, then X is split by a reversal of the CCG composition combinator, yielding:

(19) $$\{(X_1/Y : f \;\; Y/X_2 : g\}$$

where $X : h$ and $h = \lambda x.f(g(x))$. Analogous inverted combinators apply for the case of $X = X_1\backslash X_2$ and higher-order inverted composition combinators.

The categories that are introduced as a result of a split (Y in (18) and Z in (19)) are labelled via a functional mapping `cat` from semantic type T to syntactic category:

$$\texttt{cat}(T) = \left\{ \begin{array}{ll} \texttt{Atomic}(T) & \text{if } T \in \text{Figure B1} \\ \texttt{cat}(T_1)/\texttt{cat}(T_2) & \text{if } T = \langle T_1, T_2\rangle \\ \texttt{cat}(T_1)\backslash\texttt{cat}(T_2) & \text{if } T = \langle T_1, T_2\rangle \end{array} \right\}$$

which uses the `Atomic` function illustrated in Figure B1 to map semantic-type to basic CCG syntactic category.

As an example, the logical split in (17) supports two CCG category splits, one for each of the CCG application rules.

(20) $$(\mathsf{NP/N} : \lambda y.a(x, y(x)),\ \mathsf{N} : \lambda x.cookie(x))$$

(21) $$(\mathsf{N} : \lambda x.cookie(x),\ \mathsf{NP\backslash N} : \lambda y.a(x, y(x)))$$

The parse generation algorithm $\mathscr{T}$ uses the function `split` to generate all CCG category pairs that are an allowed split of an input category $\mathsf{X} : h$:

(22) $$\{(\mathsf{C_l} : m_l, \mathsf{C_r} : m_r)\} = \texttt{split}(\mathsf{X} : h),$$

and then packs a chart representation of **t** in a top-down fashion starting with a single cell entry $\mathsf{C_m} : m$ for the top node shared by all parses **t**. $\mathscr{T}$ cycles over all cell entries in increasingly small spans and populates the chart with their splits. For any cell entry $\mathsf{X} : h$ spanning more than one word $\mathscr{T}$ generates a set of pairs representing the splits of $\mathsf{X} : h$. For each split $(\mathsf{C_l} : m_l, \mathsf{C_r} : m_r)$ and every binary partition $(w_{i:k}, w_{k:j})$ of the word-span $\mathscr{T}$ creates two new cell entries in the chart: $(\mathsf{C_l} : m_l)_{i:k}$ and $(\mathsf{C_r} : m_r)_{k:j}$.

```
Input  : Sentence $[w_1, \ldots, w_n]$, top node $\mathsf{C_m} : m$
Output: Packed parse chart Ch containing t
Ch $= [\,[\{\}_1, \ldots, \{\}_n]_1,\ \ldots\ , [\{\}_1, \ldots, \{\}_n]_n\,]$
Ch$[1][n-1] = \mathsf{C_m} : m$
for $i = n, \ldots, 2;\ j = 1 \ldots (n-i)+1$ do
    for $\mathsf{X} : h \in$ Ch$[j][i]$ do
        for $(\mathsf{C_l} : m_l, \mathsf{C_r} : m_r) \in$ split$(\mathsf{X} : h)$ do
            for $k = 1, \ldots, i-1$ do
                Ch$[j][k] \leftarrow \mathsf{C_l} : m_l$
                Ch$[j+k][i-k] \leftarrow \mathsf{C_r} : m_r$
```

Algorithm 1: Generating **t** with $\mathscr{T}$.

Algorithm 1 shows how the learner uses $\mathscr{T}$ to generate a packed chart representation of **t** in the chart `Ch`. The function $\mathscr{T}$ massively overgenerates parses for any given natural language. The probabilistic parsing model (see Appendix A) is used to choose the best parse from the overgenerated set.

Appendix C
The Learning Algorithm

The parameters of the model are estimated incrementally based on one example at a time. Each example is a pair $(s, \mathbf{m})$ of an input sentence s and a set of possible meanings $\mathbf{m}$. The algorithm estimates the posterior distribution $P(t, \Theta \mid s, \mathbf{m})$: the joint distribution over possible derivations t of s and the model parameters Θ.

Since estimating this distribution exactly is intractable, we use a mean field approximation, considering a simpler family of distributions $Q(t, \Theta)$ such that Q decomposes as follows:

$$Q(t, \Theta) = Q(t)Q(\Theta)$$

We may now approximate $\mathscr{D} = P(t, \Theta \mid s, \mathbf{m})$ by finding the distribution $Q(t, \Theta)$ that minimizes the KL divergence from $\mathscr{D}$ using the Variational Bayes EM algorithm (VBEM). We adapt existing techniques for online VBEM (Beal, 2003; Hoffman, Blei, & Bach, 2010; Sato, 2001).

The algorithm iteratively performs two steps. Given an instance $(s, \mathbf{m})$, the first step ("oVBE-step") creates a packed forest representation $\mathbf{t}$ of the space of possible derivations that yield any of the observed sentence-meaning pairs (see Appendix B for details). This packed representation is used to estimate the expected number of times the expansion $a \rightarrow \gamma$ appears in the derivation space for the current example: $E_{\mathbf{t}}[a \rightarrow \gamma]$. (As above, we use the notation for syntactic expansions, but the algorithm also considers the expansion of a category into meaning representation $c \rightarrow m$ or meaning representation into wordform $m \rightarrow w$.) The second step ("oVBM-step") then computes the estimated number of times that $a \rightarrow \gamma$ appeared in any derivation of any of the observed examples. These estimates correspond to the *pseudo-counts* $n_{a \rightarrow \gamma}$ that define the Dirichlet Process defining the distribution over the expansions of a. The computation is done by extrapolating the previous value of $n_{a \rightarrow \gamma}$ with $E_{\mathbf{t}}[a \rightarrow \gamma]$. In short, the oVBE step estimates the distribution over latent derivations $\mathbf{t}$ given a (sentence, possible meanings) pair, while the oVBM step estimates the parameters of the model, represented as pseudo-counts $n_{a \rightarrow \gamma}$, given that distribution.

The full algorithm is presented in Algorithm 2. For each s and possible meaning representation $m' \in \mathbf{m}$ it uses the function $\mathscr{T}$ to generate a set of consistent parses $\mathbf{t}'$.

In the parameter update step, the training algorithm updates the pseudo-counts associated with each of the expansions $a \rightarrow \gamma$ that have been observed in any of the derivations generated for this example.

We follow Hoffman et al. (2010) in using an online update schedule that knows, prior to training, the number of training instances that will be seen N. This is a small contravention of the conditions required for the algorithm to be truly online. In practice, the value of N is used by the algorithm to weight the effects of the prior against those of the observed data and need not exactly match the amount of data to be seen. We further follow Hoffman et al. in introducing a learning rate $\eta_i = (50 + i)^{-0.8}$, where i is the iteration number. The more gradual learning rate, discussed in Section 3.3, is $\eta_i = (50 + i)^{-0.6}$.[20]

[20] For brevity, some of the figures in Section 3.3 are only presented for the 4 and 6 distractor settings. For transitives with the more gradual rate, the correct SVO category is only favored by a small margin over the alternatives, which is reflected in the CPP for individual transitives and in the CPP of the nonce transitive after a single exposure: they are

```
Input  : Corpus $D = \{(s_i, \mathbf{m}_i) | i = 1,\dots,N\}$, Function $\mathscr{T}$, Semantics to syntactic category mapping cat.
Output: Model Parameters $\{n_{a\to\gamma}\}$
for $i = 1,\dots,N$ do
    $\mathbf{t}_i = \{\}$
    for $m' \in \mathbf{m_i}$ do
        $\mathsf{C}_{\mathsf{m'}} = \mathtt{cat}(m')$
        $\mathbf{t}_i = \mathbf{t}_i \cup \mathscr{T}(s_i, \mathsf{C}_{\mathsf{m'}} : m')$
    for $a \to \gamma \in \mathbf{t}_i$ do
        oVBE-step: Compute $E_{\mathbf{t}_i}[a \to \gamma]$
        oVBM-step: Update $n_{a\to\gamma} \leftarrow (1-\eta_i)n_{a\to\gamma} + \eta_i \cdot N \cdot E_{\mathbf{t}_i}[a \to b]$
```

Algorithm 2: Learning the lexicon and the model parameters.

Appendix D
Examples of Sentence-Meaning Pairs used in the Simulation Experiments

Below are examples of sentence-meaning pairs taken from the converted dependency-annotated Eve corpus used in the simulation experiments. Predicate symbols and constants are written in the form $POS|a_f$, where POS is the symbol's part of speech, and f are additional features, where applicable (often inflectional features). However, for the purposes of this work, all symbols are viewed as non-decomposable. The parts of speech determine the semantic types, which are here omitted for brevity.

"you go ."	:	$\lambda ev.v \mid go(pro \mid you, ev)$
"you get a fly ."	:	$\lambda ev.v \mid get(pro \mid you, det \mid a(e, n \mid fly(e)), ev)$
"who are you calling ?"	:	$\lambda e.\lambda ev.aux \mid be_{PRESENT}$ $(part \mid call_{PROGRESSIVE}(pro \mid you, e, ev), ev)$
"you read about the choochoo ."	:	$\lambda ev.v \mid read_{ZERO}(pro \mid you, ev) \wedge$ $prep \mid about(det \mid the(e, n \mid choochoo(e)), ev)$

both low in absolute terms. In all other experiments and settings, the model behaves similarly as in the 4 and 6 distractor settings.