ABSTRACT

Learning a language is a daunting task, but children can learn languages efficiently and fluently within a few years, with seemingly little input data. This paradox has been a central issue of linguistic and psychological research. However, recent work on Bayesian models of language acquisition have shown two important facts: first, that cognitively-plausible computational models can learn tricky facets of language with relatively little data, and second, that such models exhibit various experimental phenomena observed in language learners. For example, Piantadosi et al’s (2008) cross-situational model of semantic acquisition surmounts two major problems in semantic acquisition (the *gavagai* problem and the subset problem), and Frank et al’s (2008) intentional model of lexicon learning exhibits mutual exclusivity (children are biased toward novel words representing novel objects) and fast-mapping (older children can learn novel words after few exposures and retain the meaning). Other computational models effectively mimic other aspects of child language learning, such as word segmentation (Goldwater 2007) or vowel categorization (Vallabha et al 2007). Such models suggest that rapid language learning is less paradoxical than previously thought; the brain is able to learn a rich language from impoverished data.

Each of the aforementioned models learns only a single component of the language, be it the lexicon, phonetic categories, or semantics. Recent work by Maurits et al (2009), though, has investigated the potential for a model that acquires both a lexicon and syntax simultaneously. This joint learning model can take advantage of bootstrapping to make the simultaneous learning task easier than the two learning tasks are separately. This paper extends this joint learner to account for ambiguity in the observed data. This inclusion of ambiguity not only creates a more accurate view of the learning problem; it also creates a new framework for incorporating world knowledge into linguistic knowledge.

The model learns an artificial language from a set of captioned pictures, where each picture shows multiple actions, and each caption describes a single action occurring in the picture. The model uses the sentence and picture pairs to learn a probabilistic context-free grammar for the language, comprising both syntactic and lexical rules. This captioned-picture design is similar to that of other models, such as Frank et al’s, and of experimental studies in adult artificial language learning, such as Amato and MacDonald (2008). From this, the model accurately learns the lexicon and syntax of a toy grammar despite receiving only ambiguous data. Furthermore, by combining world and linguistic knowledge, it produces the observed sentence processing phenomena from Amato and MacDonald’s artificial language learning experiment.
Simultaneously Learning Syntax and a Lexicon from Ambiguous Input

Gabriel Doyle

May 15, 2009

1 Introduction

Adults use language with an ease that belies its complexity. The same is true of children learning to use a language; their rapid progress and seeming ease of acquisition can make the task seem like a mere trifle. In truth, acquiring language is a daunting task, full of pitfalls. So how are children able to do it? One possibility is that the task really is too hard, and that children must rely on an innate cognitive framework that is activated by encountering language. The child can then take the observed language data and use it to flesh out the innate language skeleton. Another possibility is that the task is hard, but a child’s cognitive abilities are up to the task. More general cognitive abilities such as statistical learning are sufficient, in this view, to construct a language from the input a child receives. In this paper, we follow the second possibility and see how far statistical learning can take us.

We propose a Bayesian joint-learning model that learns the syntax and lexicon of a language using general statistical learning methods without an innate linguistic framework. The model is a joint learner, in that it learns the syntax and lexicon simultaneously. This allows for immediate bootstrapping; each new bit of information about the syntax is immediately used to help learn the lexicon, and vice versa. The model is also cross-situational, in that it can use knowledge gained from past linguistic experience to inform its judgments about new experiences. This allows the model to compensate for the lack of negative evidence that a language learner receives and to make effective use of limited input. These two traits allow the model to overcome the *gavagai* problem and other difficulties in acquisition and learn a highly accurate syntax and lexicon. Furthermore, the generative process underlying the model applies not only to the acquisition task, but to other language tasks as well. In the current study, the model is applied to sentence processing tasks in order to combine syntactic and lexical information with world knowledge without appealing to a lexicalized grammar. We show that this application makes processing predictions that agree with reading data from adults learning an artificial language.

The paper proceeds as follows. Section 2 situates the model within the debate about the methods of language acquisition. This section establishes some of the difficulties in acquisition, and discusses what models can and have revealed about the acquisition problem. The specific acquisition issues that the present model focuses on are discussed in Section 3. Section 4 is the heart of the paper, containing the acquisition situation as seen by the model, the model’s cognitive plausibility, and the mathematics underlying the statistical learning methods. Section 5 looks at four experiments and how they address the problems set out in Section 2. Section 6 discusses the results and some open questions about the model. Section 7 lays out the infrastructure for forthcoming extensions of this model, and Section 8 concludes.

2 Background

2.1 Language Learning and its Hurdles

Cognition is a many-splendored thing; humans learn to locate objects using stereoscopic vision, to construct and manipulate mental maps, and to infer cause-and-effect relationships in the world. But perhaps no cognitive ability is held in as high esteem as the ability to learn to use language. Language usage is one of the most obvious distinctions between humans and other animals, and perhaps one of the most complicated abilities to acquire. But how does this acquisition take place, and what mechanisms does it use? Can language acquisition be explained by general cognitive mechanisms, or is it a task that is insurmountable without innate, language-specific\(^1\) cognitive scaffolding? This question has divided language researchers into two camps: the nativists, who argue that language is too complex to acquire without a cognitive helping hand, and the empiricists, who argue that the cognitive abilities used in other aspects of cognition are sufficient to acquire language as well. For much of its history, the debate has been short on evidence and long on assertion. However, advances in the modelling of lan-

\(^1\)“Language-specific” in this paper means specific to language learning and processing (as opposed to vision, orientation, or other cognitive processes), but not restricted to a specific language.
language acquisition have helped to clarify the picture by revealing the power of statistical learning and general cognitive mechanisms in the linguistic sphere, suggesting that the need for language-specific mechanisms has been overstated.

### 2.1.1 The Logical Problem of Language Acquisition

The nativist view is, in a key sense, the pessimistic view of language acquisition, as it supposes that general cognitive mechanisms are weak, the language is unapproachably complex, and that the language available to a child is impoverished. These last two points combine to form Chomsky’s argument from the poverty of the stimulus (Chomsky, 1965, 1975). This argument has three components. First, the language a child hears is noisy; it contains factual and grammatical errors. Second, any subset of the possible utterances in a language will not fully specify the language’s underlying structure. This is a result of the “language bottleneck”: a language can specify an infinite number of sentences, but a child can only hear a small finite set of all possible sentences. Third, the input data contains little negative evidence for the child to learn what is ungrammatical. Adults do not intentionally produce ungrammatical sentences to show children what is ungrammatical, they do not often correct children’s or their own ungrammatical utterances, and children seem to ignore corrections of their usage when they do receive them (McNeill, 1966; Brown & Hanlon, 1970). When the poverty of the stimulus argument is combined with the claim that general cognitive processes would be too weak to learn abstract language, even if the input were richer, these problems are referred to with the blanket term the logical problem of language acquisition (Baker & McCarthy, 1991).

If the language acquisition task is so difficult, how is language ever acquired? The nativist answer is that language is learnable only because of the innate Language Acquisition Device (LAD), which relies in turn on the grammatical infrastructure of Universal Grammar (UG). UG and the LAD simplify the acquisition problem in a few ways. UG constrains human language, since each language must be consistent with UG in order to be learned and passed on through the generations. Thus UG limits the search space of possible languages the child must consider. UG also gives the child an innate knowledge of the central abstract notions in language, which nativists argue cannot be learned. The LAD contributes as well, as its language-specific mechanisms are able to extract the structure of any language consistent with UG. Thus the formerly impossible problem becomes solvable through specialized equipment; if language acquisition were a mountain, LAD and UG would be the crampons and rope that make it summable.

Various formulations of UG and the LAD have been proposed to explain how these ideas can simplify the logical problem of language acquisition. One popular form is the principles and parameters model (Hyams, 1986; Roeper & Williams, 1987; Hyams & Wexler, 1993), in which UG is likened to a vast array of switches. Each switch, which corresponds to some aspect of a language, has some small set of possible settings. For instance, one switch might be headedness, which may be set either to left- or right-headed. If the child is exposed to English, the left-headedness setting is chosen, while if the child is exposed to Japanese, the right-headedness setting is chosen. This parameter-setting model simplifies the acquisition problem in two major ways. First, it clearly delineates the search space of possible language-forms the child will consider, especially since some parameters are argued to be correlated (Greenberg, 1963; Hyams, 1987). Second, reducing the problem to parameter-setting allows certain features of the input to function as “triggers”, which the child can use to quickly set some parameters from very little input (Wexler & Manzini, 1987; Dresher & Kaye, 1990; Dresher, 1999). Quickly setting parameters simplifies the acquisition problem even further, as each set parameter reduces the number of language hypotheses the child must entertain.

Three additional types of evidence are cited in favor of the nativist view, in addition to the logical problem of language acquisition. The first relates to the paradox that abstract language is uniquely and universally human. Non-human animals, even other moderately intelligent primates, appear to be unable to learn more than the rudiments of human language (Allen & Gardner, 1969), suggesting that whatever cognitive processes allow humans to learn language must be restricted to humans. Yet humans over almost the whole range of human intelligence are able to learn a language natively. Having an innate LAD could explain both of these; all humans, regardless of other cognitive abilities, have the LAD, accounting for the universality, and only humans have it, accounting for the uniqueness. On the other side of the coin, Specific Language Impairment (SLI) has been cited as evidence that language abilities may be specifically afflicted while other cognitive processes remain unharmed. This can be explained with the LAD as well if SLI involves impairment of the LAD but not of general cognitive abilities. One last set of arguments looks at the ability to create and extend the structure of a language. Bickerton (1984) argues that the rapid creolization of pidgins is best explained if children have an innate grammatical infrastructure with which to mold the unstructured pidgin. Similarly, the rapid creation of complex sign languages from simple home sign systems, as happened with Nicaraguan Sign Language, has been explained as the result of combining simple input with innate grammatical knowledge to create a complex language (Kegl, Senghas, & Coppola, 1999; Kegl, 2002).
2.1.2 Data-Driven Accounts

Empiricists reject one or more of the suppositions that compose the logical problem of language acquisition. Empiricists contend that the same general-purpose mechanisms that drive other cognitive processes are generally sufficient to acquire language as well. Some language-specific processes may be used as well (Elman et al., 1996), but the driving forces behind acquisition are general mechanisms, such as statistical learning. The key difference is that empiricists argue that there is no clear need for extensive innate cognitive scaffolding in language acquisition. This argument can take two forms: that language acquisition is not as insurmountable a task as the nativists claim and/or that general mechanisms are more powerful than originally thought. Support for the first of these arguments comes from empirical studies of language data, while work on models of language acquisition support the second argument.

One major result in the empiricists’ favor is that the poverty of the stimulus seems to have been overstated (Elman, 2003; Pullum & Scholz, 2002). Newport, Gleitman, and Gleitman (1977) described child-directed speech as generally following the conventions of “motherese”, a sort of simplification of the language being learned. Caregivers tend to speak about objects in the child’s attentional focus, to stay on the same topic and repeat or restate previous sentences. Fernald et al. (1989) showed that child-directed speech is usually slow and more grammatically correct than adult-directed speech. Aspects such as these reduce the noise in the input, giving the learner more useful input data than was originally thought. However, using motherese in child-directed speech is not universal (Lieven, 1994), so it remains a priori plausible that the input language children receive is at least as noisy as adult-directed speech.

This is a core problem for the logical problem of language acquisition; much of it boils down to assertions, that although fitting with common sense, are not well-supported with experimental evidence. The poverty of the stimulus argument assumes that speech data is too noisy to construct a grammar, but does not clearly quantify the amount of noise in the input, nor does it investigate how much noise is too much for successful acquisition. Arguments citing the lack of direct negative evidence assume that only direct negative evidence is useful. However, recent work has shown that even in the absence of direct negative evidence, Bayesian reasoning creates indirect negative evidence (e.g., Tenenbaum, 1999; Plantadosi, Goodman, Ellis, & Tenenbaum, 2008). As a result, determining the extent, if any, of innate language-specific cognitive mechanisms depends crucially on the extent to which the nativist assertions are valid.

Innate language-specific structure is only necessary if general cognitive mechanisms are truly insufficient for the acquisition task. Thus it is necessary to determine what the general mechanisms are capable of learning. Models of language acquisition are essential tools for this test, as they are controlled experiments to determine whether a general mechanism can overcome specific hurdles in language acquisition.

2.1.3 Hurdles in Language Learning

We consider four such hurdles in language learning, and discuss modelling evidence suggesting how these hurdles may be overcome through general cognitive mechanisms.

The gavagai problem The first problem is the GAVAGAI PROBLEM, also known as the PROBLEM OF REFERENTIAL UNCERTAINTY: that an uttered word or sentence can conceivably refer to any of an infinite number of meanings. Consider the case, presented by Quine (1960), of a learner hearing a native speaker of a language utter the word gavagai while looking at a white rabbit. While the obvious meaning seems to be “rabbit”, there are other meanings that would be valid as well, such as “white” or “quadruped”. Complicating matters further is the fact that the utterance gavagai may not be not related to anything apparent in the present situation. The speaker may be indicating some mental state, such as hunger or irritation, that is completely unrelated to the rabbit. The problem, then, is how one can determine the intended referent or meaning of an utterance, given that this intention cannot be made perfectly evident to the learner.

Frank, Goodman, and Tenenbaum (2007); Yu and Smith (2008) use models of lexical acquisition to show that the gavagai problem can be mitigated by cross-situational learning. Cross-situational learning arises when a learner is able to use information gleaned from one situation to inform judgments in another. This can work around the gavagai problem; if the learner hears gavagai around not only rabbits, but also elephants, bats, and beagles, the learner can deduce that gavagai refers to more than just rabbits, and might propose that it likely means something like “mammal”. This definition could then be modified in light of additional information; if the learner hears gavagai in the presence of a sparrow, for instance, the definition could be revised to the more general “animal”, since a sparrow is not a mammal. Overgeneralization of the meaning (e.g., thinking that gavagai means “animal” when it really means “rabbit”) is held in check by the size principle (Tenenbaum, 1999), discussed below.

Negative evidence and the subset problem It is unclear exactly how much negative evidence children encounter when learning a language, but the negative evidence is clearly limited and is not well-attended to (McNeill, 1966; Brown & Hanlon, 1970). For instance, children do not seem to be explicitly told that Verb-Subject-Object word order is ungrammatical in English. Yet somehow children learn to avoid what is ungrammatical in their language and by adulthood codify this avoidance into an abstract condition of ungrammatical-
ity. This seeming contradiction of learning what is unacceptable without being directly informed has been a central part of the poverty of the stimulus argument. A similar contradiction emerges in semantics; without negative evidence, a learner could stick with a weak semantic hypothesis. Quantifiers are especially susceptible to this problem; every time that a sentence like “all ducks have feathers” is true, the weaker statement “some ducks have feathers” is also true. Without negative evidence, it seems a learner should be equally likely to learn that all means “some” as to learn the true meaning of all.

This is referred to as the subset problem.

However, the preference for the correct grammar and semantics over weak plausible ones can be explained by general statistical principles. The weak hypothesis that all means “some”, or that VSO word order is acceptable in English, does not fit the data well and thus is dispreferred (Piantadosi et al., 2008). This dispreference arises from probabilistic learning. Suppose a model is choosing between a weak grammar that allows SVO and VSO word order and a strong grammar that only allows SVO. The weak grammar would be expected to produce both SVO and VSO sentences, while the strong grammar would only produce SVO sentences. Since the English data contains almost entirely SVO sentences, it is a more likely result of the strong grammar than the weak grammar, and so the strong grammar will be preferred by Bayesian inference (Section 4.4). This result is called the size principle (Tenenbaum, 1999) and suggests that even without negative evidence, a learner can avoid weak grammars using statistical principles, without relying on a language-specific mechanism.

Mutual exclusivity If a child hears a novel word in the presence of a familiar object and a novel object, the child will infer that the novel word refers to the novel object (Markman & Wachtel, 1988; Byers-Heinlein & Werker, in press). This phenomenon is referred to as mutual exclusivity or disambiguation. Many explanations have been proposed for this. Clark (1987) has argued for a social-pragmatic constraint that different words have different meanings. Diesendruck and Markson (2001) have argued for a similar pragmatic constraint, although theirs is based on the assumption that one would use a known word to refer to a known object. Markman and Wachtel (1988) argue that this comes from a cognitive mutual exclusivity constraint that makes the child assign no more than one word to each object. One last proposal is the Novel-Name Nameless Category assumption of Mervis and Bertrand (1994), which claims that children want to find a name for every object.

The first two proposals rely on pragmatic constraints to get the exclusivity effect, while the last two rely on specialized cognitive rules. Acquisition models offer a potentially better explanation; Frank et al. (2007) show that mutual exclusivity can arise naturally from a Bayesian quest to find the lexical assignments that best explain the observed data, similar to the subset problem described above. This avoids the need for any special principles to impose mutual exclusivity.

Bootstrapping Bootstrapping refers to the use of one piece of linguistic information to aid in learning another. The two most commonly discussed forms of bootstrapping are semantic and syntactic bootstrapping. Semantic bootstrapping refers to using knowledge of the meanings and argument structures of a word (usually a verb) to predict its syntax (Grimshaw, 1981). For instance, knowing that the action of hitting involves both a hitter and a hittee makes it likely that it will occur with transitive syntax. Syntactic bootstrapping refers to using the syntactic (and morphological) context a word appears in predict its meaning (Landau & Gleitman, 1985). For instance, a child hearing “look at the zog” can infer that zug is likely a noun. Bootstrapping is clearly a powerful mechanism in language acquisition, but it leads to a chicken-or-egg problem. Semantic bootstrapping requires syntactic knowledge, and syntactic bootstrapping requires semantic knowledge. Certainly, if knowledge of one is given, bootstrapping can be effective, but how is the first piece of knowledge gained?

Computational models have offered two plausible mechanisms. The first is cross-situational learning, where social cues are used to bootstrap lexical meanings (Frank et al., 2007), which can then be used for syntactic bootstrapping. Another mechanism is joint learning, where syntax and semantics are being learned simultaneously. Joint learning allows for bootstrapping from incomplete knowledge; statistical trends in the syntax constrain the lexicon and vice versa. This approach is taken by Maurits, Perfors, and Navarro (2009) as well as the model presented in this paper. In joint learning models, bootstrapping emerges naturally as a consequence of the joint learning problem.

These four examples show that modelling acquisition can reveal new explanations for well-known problems. The next section delves into a brief history of acquisition models, what they have revealed, and how their purposes have changed.

2.2 Modelling Language Acquisition

Models using statistical data to learn linguistic structure have been around for at least fifty years, starting with the word segmentation model of Harris (1954). The exact goals of these models have varied, but the general purpose has been to investigate what conditions are nec-

---

2Of course, both grammars must allow the other word orders, with very low probability, to account for sentences that are topocalized or left-dislocated and the like. We ignore these as their probability mass is small.

3Explaining mutual exclusivity without resorting to specialized social-pragmatic constraints is especially important as this ability seems present in at least one dog as well (Kaminski, Call, & Fischer, 2004).
statistical learning can play a significant role. That language acquisition need not be entirely innate; language (e.g., Saffran, Aslin, & Newport, 1996), suggested children are sensitive to statistical information in language data available to children had significant statistical the key point is that these early models showed that language contained potentially useful patterns for language acquisition. The newer models have begun to exploit these statistical patterns to learn language in a cognitively plausible method, with less language-specific infrastructure.

Perhaps the most interesting class of these new models are the Bayesian-optimal models. These models propose that "human cognition may be explicable in rational probabilistic terms and that, in core domains, human cognition approaches an optimal level of performance." (Chater, Tenenbaum, & Yuille, 2006, p. 289) This seems a slightly odd proposal, in light of the many fallacies observed in human probabilistic reasoning — for example, the continued success of lotteries, despite their heavy odds-weighting against the ticket-buyers, would seem a heavy blow to this claim. Fallacies are commonplace in a variety of economic and ethical decision-making tasks (Kahneman & Tversky, 2000), and occur even in very simple probabilistic tasks (Tversky & Kahneman, 1983; Shafir, Smith, & Osherson, 1990). How then could humans be considered to be near-optimal in any probabilistic task?

There are two answers to this. The first is that human brains are able to perform a variety of other difficult mathematical operations subconsciously. Hunters using projectiles, whether spears, arrows, or guns, must perform complex calculations to account for distance, wind speed, gravity, and other factors if they are to hit their targets. Yet hunters had been successful for thousands of years before the calculus that could solve such problems was invented. Similarly, the human auditory and visual systems rely on Fourier analysis, which is an imposing branch of mathematics. The lack of conscious knowledge of an area of mathematics does not preclude its use in cognitive processes. In fact, Chater et al. (2006) note that optimality may be more closely approached in more unconscious and specific processes, such as vision, that humans have more extensive exposure to, while processes that require explicit mathematical computation may be further from optimality.

The second answer is empirical; optimal models are good fits for human behavior. Observed human performance in many cognitive processes fits with the predictions of optimal models, and such models may explain seemingly sub-optimal performance in certain tasks. Vision presents perhaps the best evidence for this, as it has been extensively modelled and work in vision has laid important groundwork for modelling other cognitive processes (including Marr’s three levels of computational explanation). Weiss, Simoncelli, and Adelson (2002) showed that an optimal model, assuming noisy ob-

2.2.2 Bayesian-Optimal Models

With these early models in showing that useful statistical patterns exist in the input, more sophisticated models arose, delving deeper into the nature of learning and learnability in human cognition. The early models showed that language contained potentially useful patterns for language acquisition. The newer models have begun to exploit these statistical patterns to learn language in a cognitively plausible method, with less language-specific infrastructure.

Perhaps the most interesting class of these new models are the Bayesian-optimal models. These models propose that "human cognition may be explicable in rational probabilistic terms and that, in core domains, human cognition approaches an optimal level of performance." (Chater, Tenenbaum, & Yuille, 2006, p. 289) This seems a slightly odd proposal, in light of the many fallacies observed in human probabilistic reasoning — for example, the continued success of lotteries, despite their heavy odds-weighting against the ticket-buyers, would seem a heavy blow to this claim. Fallacies are commonplace in a variety of economic and ethical decision-making tasks (Kahneman & Tversky, 2000), and occur even in very simple probabilistic tasks (Tversky & Kahneman, 1983; Shafir, Smith, & Osherson, 1990). How then could humans be considered to be near-optimal in any probabilistic task?

There are two answers to this. The first is that human brains are able to perform a variety of other difficult mathematical operations subconsciously. Hunters using projectiles, whether spears, arrows, or guns, must perform complex calculations to account for distance, wind speed, gravity, and other factors if they are to hit their targets. Yet hunters had been successful for thousands of years before the calculus that could solve such problems was invented. Similarly, the human auditory and visual systems rely on Fourier analysis, which is an imposing branch of mathematics. The lack of conscious knowledge of an area of mathematics does not preclude its use in cognitive processes. In fact, Chater et al. (2006) note that optimality may be more closely approached in more unconscious and specific processes, such as vision, that humans have more extensive exposure to, while processes that require explicit mathematical computation may be further from optimality.

The second answer is empirical; optimal models are good fits for human behavior. Observed human performance in many cognitive processes fits with the predictions of optimal models, and such models may explain seemingly sub-optimal performance in certain tasks. Vision presents perhaps the best evidence for this, as it has been extensively modelled and work in vision has laid important groundwork for modelling other cognitive processes (including Marr’s three levels of computational explanation). Weiss, Simoncelli, and Adelson (2002) showed that an optimal model, assuming noisy ob-
servation and a prior belief favoring slower motion, predicted certain motion illusions reported in human vision. Illusions are not necessarily evidence against optimality. This suggests that shortcomings in other cognitive processes are not, in and of themselves, compelling evidence of human sub-optimality.

In language acquisition, too, optimal models can be effective at improving our understanding of the underlying mental processes. We look at four recent models, each working in a separate area of language acquisition, to see what these optimal models can reveal.

Word segmentation Goldwater (2007) proposed a Bayesian model of word segmentation that reads in a stream of phonemes and decides where to draw word boundaries. The model is generative, in that it specifies the process by which it assumes that the phonemes it sees are generated. The generative process has two stages. First, a lexicon is generated by choosing various phoneme strings to be the words of the language. This lexicon remains hidden from the learner. Then the observed stream of phonemes is generated by choosing words from the lexicon, concatenating them, and erasing the word boundaries. The model, encountering a stream of phonemes, uses Bayesian inference to determine which lexicons are likely sources for the observed phonemes. The Bayesian inference uses a minimum-description length prior to favor simpler lexicons.

Goldwater tests a version of this model whose generative process supposes that the phoneme strings are generated by a bigram probability distribution. This model, given a corpus of transcribed child-parent conversation from the CHILDES database with the word boundaries removed, identifies more than 83% of the true word boundaries (recall), and 92% of the identified boundaries are true boundaries (precision). These boundaries are identified using only statistical information from the phonemes themselves; the transcribed corpus contains no information about prosody or intra-phoneme variability. The model can also be used to learn morphology, correctly decomposing over 70% of the verbs in the Wall Street Journal corpus into their stems and suffixes. This shows that statistical learning can drive word and morpheme segmentation.

Word learning Frank et al. (2007) proposed a generative Bayesian model for cross-situational word learning using social cues. The generative process underlying this model starts with a set of objects visible in a scene. Words are generated by the speaker choosing an object or objects to name (the *intent*) and then drawing their names from the lexicon and uttering them. The choice of intent also generates social cues; the speaker might look at the objects being named, or touch them, or point to them. The input data coming into the model is a corpus of utterances and the social cues valid during each utterance. The model then uses Bayesian inference to infer the lexicon using the cross-situational data and social cues.

Frank et al. test the model on videotaped child-parent interactions from the CHILDES database, and show that the model displays three important characteristics of child language acquisition. The primary characteristic is that the model exhibits mutual exclusivity: novel words are assumed to map to novel objects. This arises directly from the model’s preference for Bayesian-optimal lexicons, as discussed in Section 2.1.3. The reason for this can be seen by considering the case where a child is presented with a ball and a novel object, and hears the word *daz*. Probabilistically speaking, it is unlikely that the child has always heard the ball being referred to as a ball if *daz* is an acceptable word for it, so a Bayesian model will find the mapping of *daz* to the novel object more likely. This can explain mutual exclusivity as a purely statistical phenomenon, rather than as a result of language-specific or social-pragmatic principles. Fast-mapping (learning and retaining a novel word after few exposures) and word learning from social cues, both attested in children’s acquisition, also arise as Bayesian-optimal behavior in this model.

Quantifier acquisition Piantadosi et al. (2008) proposed a cross-situational Bayesian model for learning the compositional semantics of quantifiers. This model requires a lexicon to have already been learned for the nouns and verbs, although this could in principle be learned from another Bayesian model, such as Frank et al’s. The model uses lambda calculus to represent the semantics and Combinatory Categorical Grammar for the syntax because these two frameworks fit together well, although other representations could be used. The goal is to learn an effective semantic grammar, a function from words in the lexicon to CCG syntactic types and lambda expressions. As with the preceding models, this is done through Bayesian inference. The input is a set of sentences uttered in varying contexts, where each context is a set of propositions that are true in the world. The Bayesian inference is then used to infer a grammar from the utterance-context pairs, with a prior belief favoring simpler grammars. This model also directly accounts for noise in the data by setting aside probability mass for the potential case where the utterance does not match the context.

Piantadosi et al test the model with a toy version of English, and show that the model’s grammar reaches ceiling accuracy with a moderately-sized corpus. Furthermore, the model captures and potentially explains some acquisition phenomena. First, the model learns the correct compositional semantics despite the *gavagai* problem; a range of propositions are true for each context, yet the grammar uses the cross-situational nature of the input to determine which propositions each utterance refers to. Second, the model overcomes the subset problem.

---

4These contexts are similar to the pictures in our model; see Section 4.
A joint-learning model. The models above are limited because each learns one component of language in isolation from the others, or requires some components to already be known to the learner. The Goldwater model learns word segmentation without any knowledge of the words’ meanings; the Frank et al. model acquires its lexicon independent of the syntax and semantics of the language and requires the input to be segmented; the Pi-antadosi et al. semantic model requires the lexicon to already have been learnt before it acquires compositional semantics. The next question is then investigating how the interaction of these components can affect their acquisition, especially in light of the potential for bootstrapping. This line of research has been kicked off by Maurits et al. (2009), who looked at the joint acquisition of word order and lexicon, using a Bayesian-optimal model.

The Maurits et al. model takes as its input a series of simple transitive sentences, each uttered in its own unambiguous context, and learns both the probability distribution over word order and the distribution over word referents. The model uses the following generative model; first, a context is chosen from a multinomial distribution over possible situations in the world. The context consists of a relation r that holds between a pair of objects (o1, o2). The context “Cat eats mouse”, for instance, would be represented by EAT(CAT, MOUSE). The lexicon takes the form of mappings from meanings (relations or objects) to observed words. These mappings are referred to as the naming distributions λx, where x is a relation or object. Each λx is a well-defined probability distribution; if V is the entire vocabulary of the language, then \( \sum_{w \in V} \lambda_x(w) = 1 \).

This model succeeds on two fronts. First, it successfully learns both the lexicon and word order. Word order is learned accurately even with very small datasets, and word order is learned reasonably accurately (about 70% correct with a large dataset). Secondly, the lexicon is learned more accurately and faster with their joint-acquisition model than with a baseline model that does not take word order into account while learning the lexicon. This suggests that learning two aspects of language simultaneously can make both learning problems easier.

2.3 Explanations at Different Levels

These four models show some of the advantages of Bayesian-optimal models. They can learn difficult aspects of language using little to no language-specific cognitive processes. Moreover, they exhibit various experimental phenomena observed in human language learners. These models suggest that language-specific cognitive mechanisms are less necessary than originally thought, and that general cognitive processes such as statistical learning are more powerful and more explanatory in the linguistic domain than expected. But are such computational models really appropriate simulations of human cognition?

Models come at a price; all models require some simplification of the task being studied. This is an unavoidable consequence of modelling a phenomenon as complex as language acquisition, but it raises an important question. What can be inferred from the results of the model, if the model does not map directly onto an infant? The answer lies in Marr (1982, pg. 25)’s levels of information processing, which gives three levels at which a phenomenon can be modelled.

The first, and most general, is the computational level. This level addresses the question, “What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?” The second is the algorithmic level, and addresses the question “How can this computational theory be implemented?” The last, and most specific, is the implementational level, which addresses the question “How can the representation and algorithm be realized physically?” In rough terms, these three levels correspond to the mind, brain, and neurons respectively.

These levels are only weakly related. An explanation at the computational level can be realized in many different ways at the algorithmic and implementational levels. For instance, to solve the task of lifting a weight, human arms and robot arms use similar methods at the computational level, applying force to the fingers to grasp the weight and using the elbow and shoulder as fulcrums to lift the weight as with a lever, but they rely on strikingly different hardware, with the former using muscles and a nervous system while the latter uses pistons and electrical wiring. Likewise, similar explanations at the implementational level may apply to vastly different tasks at the computational level. The neural structure of the human brain (as an implementational level) can be used for a dizzying array of tasks at the computational level, and likewise a single programming language can be made to perform virtually any calculation.

Although there have been models of language acquisition at the algorithmic and implementational levels, notably the connectionist models that directly accounted for the neural architecture, recent models have focused instead on the computational level. There are two major reasons for this. First, the issue being investigated by these models is how statistical learning fares as an acquisition strategy. Because this is a question of computational strategy, it is best answered at the computational level. Secondly, the algorithms that the human brain uses
and their physical implementation remain murky — not for lack of research, but rather as a result of the tremendously difficult nature of the identification task. As such, it is unclear how to judge the goodness of acquisition models at the algorithmic and implementational levels.

Both the Bayesian-optimal models of the preceding section and the model proposed in this paper are computational-level proposals about the methods of language acquisition. We are interested in establishing the learning task and looking at how the different sources of available information can be used and can interact to solve the learning task. These issues exist independently of the algorithmic and implementational levels; any algorithm can be used so long as it solves the problem. This is a crucial point about computational models of human cognitive activity, since computational models are generally implemented on computers, whose architecture is quite unlike the brain’s. As a result, all of these models are inaccurate at the implementational level, but this does not impugn the results at the computational level.

There is one condition on models that represent human cognition; in the end, the computational-level model must be implementable with a cognitively-plausible algorithm. It is not clear what the set of cognitively-plausible algorithms are, and thus this question cannot be answered definitively, but Section 4.3 argues that this model can likely be implemented with a cognitively-plausible algorithm.

3 Purpose

In light of the previous modelling work, the immediate question for any new model is “What can this model demonstrate that previous models have not?” The model proposed in this paper is foremost an extension of Bayesian models of language acquisition to a new and more complicated learning task. Like Maurits et al. (2009)’s model, the current model does joint acquisition of syntax and lexicon. Unlike Maurits et al.’s model, which assumes that utterances are uttered in unambiguous contexts, our model goes one step further and allows for ambiguous utterance contexts. This improves the plausibility of the learning task, and the model’s successful acquisition of the syntax and lexicon shows that joint acquisition and cross-situational learning can overcome such ambiguous contexts. This also shows that bootstrapping, in the form of joint-learning, remains possible under referential ambiguity. This shows evidence for the strength of statistical learning and further argues against the need for language-specific cognitive structure.

The model also proposes a novel way of storing linguistic knowledge (i.e., the grammar learned by the model) and world knowledge (i.e., the observed contexts of the utterances) separately, while still allowing the world knowledge to influence linguistic judgments. This results in a grammar that can use world knowledge without being lexicalized or incorporating any additional structure. The world knowledge is stored completely outside the grammar and is factored in using a simple product-of-experts calculation, as shown in Section 4.6. This can be useful in grammar induction research, since it offers a way to account for world knowledge while keeping the grammar from becoming overly complex.

Lastly, this separation of linguistic and world knowledge allows us to make testable predictions about the nature of language acquisition. It predicts that increased exposure to world knowledge, even in the absence of increased linguistic exposure, can improve one’s psycholinguistic expectations. This comes from the influence of word knowledge in the product-of-experts calculation; better knowledge of the likelihood of events in the world improves the model’s ability to predict sentence completions. This combination of grammar and world knowledge can be investigated at various points in the time course of learning and tested against empirical data from online sentence processing in artificial language learning experiments; in fact, we show that the model’s expectations about sentence completions agree with learner data in an artificial language learning experiment (Amato & MacDonald, 2008). The following subsections discuss each of these points in more detail.

3.1 Robustness to Ambiguity

The first advance of this model is its robustness to the ambiguity inherent in the language learning process. The joint acquisition model of Maurits et al. (2009) successfully acquires the grammar and lexicon of a language, but it does so in an unambiguous world, where it is clear to the learner what each sentence is intended to mean. This is a significant simplification of the learning task an infant is presented with; real learners must overcome substantial ambiguity in the input. Although children can and do use social and attentional cues to disambiguate the referents for an utterance (Hollich, Hirsh-Pasek, & Golinkoff, 2000), some ambiguity will still remain about the intended meaning of an utterances. Because of this remnant ambiguity, a convincing model of language acquisition must be able to account for the ambiguity of sentences uttered in the world. The model in this paper can learn a language despite this ambiguity, showing that the statistical learning is sufficiently powerful to overcome ambiguous input when learning both the meaning of words and the structure of utterances. The model observes its training utterances along with pictures, which describe a set of propositions that are true in the context of the utterance. Thus the meaning of an utterance is ambiguous, and the model must learn to disambiguate the meanings for itself.

3.2 Separation of Linguistic and World Knowledge

This model also proposes a means of using world knowledge to inform linguistic judgments while storing
the two types of knowledge separately. In these sentences, knowledge about likely events in the world affects the building of syntactic expectations. For instance, Trueswell, Tanenhaus, and Garnsey (1994) showed that readers were slower in reading reduced relative clauses with an animate NP (the defendant examined by the lawyer) than an inanimate NP (the evidence examined by the lawyer), because of the initially plausible analysis that the animate NP was the subject in a main verb reading of the sentence (e.g., the defendant examined the evidence). Similarly, McRae, Spivey-Knowlton, and Tanenhaus (1998) showed that participants in a self-paced reading time task were slower at reading by the detective in the NP the cop arrested by the detective than in the NP the crook arrested by the detective, due to the cop being a much more common agent than patient for arrested and thus biasing the readers toward the main-verb reading. Clearly, people use their world knowledge to build their expectations about sentence continuations.

So how can this knowledge about the world be incorporated into sentence processing? Our model incorporates world knowledge as contexts that influence what utterances will be issued. This is an essential component for the learning process; without the contexts, there would be no way for the model to learn the meanings of words. After the grammar is learned, the model structure offers a framework by which world knowledge can be used to influence linguistic judgments, through a product-of-experts calculation. As discussed in Section 4.6, by placing a distribution on the truth-values of possible states of the world, the generative process in the model allows for the calculation of the probabilities of possible sentence completions, taking account of world knowledge. This setup can be used to investigate the effects of common versus uncommon event structures on reading times.

This also suggests a new method of using world knowledge in automated parsers. Currently, to get world knowledge into a parser, a lexicalized grammar has to be used. However, our model compartmentalizes world knowledge separately from linguistic knowledge before recombining them in processing. This may be a reasonable approach for an automated parser; the parser can use its linguistic knowledge to come up with possible parses for a sentence, and then use world knowledge to weight these possibilities. This would be akin to the way a speech-recognition system combining a language model and an acoustic model to determine the probability of each possible interpretation. The product-of-experts setup can do just this.

3.3 Learning Unimodally

Furthermore, if knowledge about the world affects sentence processing, we should expect that a model can extract useful information not only from paired linguistic and non-linguistic data (such as the sentence-picture pairs in this model), but also from unpaired data. For instance, hearing an utterance in the absence of a picture can still yield usable information about its structure, according to syntactic bootstrapping; hearing this is a zug strongly suggests that zug is a noun, even though its exact meaning may be unclear. Similarly, observations about the world, even if unaccompanied by language, can improve knowledge about which states of the world are likely and which are not — a child can learn that cops tend to arrest criminals, rather than the other way around, solely from world observations. Our model captures this ability to learn from unpaired data as well as paired data; further exposure to pictures, unaccompanied by sentences, will improve sentence processing by making more common events more predictable.

3.4 Online Expectation Generation

Lastly, this model can account for the observed ability of adult learners of an artificial language to quickly use world knowledge in online sentence processing. Amato and MacDonald (2008) showed that language learners are able to learn second-order dependencies, such as likely objects given a subject-verb pairing, in an artificial language learning paradigm. Furthermore, they use this information in sentence processing. Participants were tested in a VSO artificial language where each subject-verb pair had one preferred object that occurred more often than other objects (i.e., the red monster breathed fire on the blue monster more than on any other monster). The preferred object was different for different subject-verb pairs. After training in the language, learners read a sentence with the preferred object or a dispreferred object, and reading times were significantly greater for dispreferred objects. We show that this observation can be explained through expectations generated in our model by combining the observed world knowledge and the learned grammatical and lexical knowledge.

Various psycholinguistic phenomena are tied to world knowledge; for instance, knowledge of the common agents and patients for a verb affect reading times of a reduced relative clause.

4 The Model

4.1 The Situation and its Simplification

Our model is concerned with the acquisition of a lexicon and a grammar from sentences uttered in varying situations. To model this, we must consider a setup where the learner has the ability to learn cross-situationally and is presented with the set of possible referents from each situation. We propose the following setup for the learning task as a reasonable facsimile of the real learning task. The model takes as its training data a set of sentences $S = \{S_1, \ldots, S_N\}$, with each sentence $S_i$ paired with a picture $\pi_i$. The sentences contain no information about their meanings or syntactic structures; each is simply a string of lexical items. Each picture is a set of proposi-
4.2.1 The Grammar

4.2 Model Components

The sentences $S$ are generated with a probabilistic context-free grammar (PCFG). A PCFG has five components:

1. A start symbol $N_0$
2. A set of non-terminal nodes $\{N_i\}$
3. A set of terminal nodes $\{w_i\}$
4. A set of productions $\{N_j \rightarrow R_k\}$, where $R_k$ is a sequence of terminals and non-terminals
5. A probability function $p$ over productions such that $\sum_k p(N_j \rightarrow R_k) = 1$ for all $j$.

A PCFG is a simple representation of phrase-structure that can be used to generate syntactic trees. One starts with the start symbol $N_0$ and successively re-writes each non-terminal node using one of the productions, until all non-terminals have been re-written as terminals. In most linguistic contexts, the terminal nodes are lexical items, while the non-terminals are syntactic categories.

The probability function $p$ determines the probability of a given non-terminal re-writing as a given sequence of terminals and non-terminals. Since each non-terminal must re-write, the probability over all possible re-write sequences for a given non-terminal must sum to 1. The probability of a tree is the product of the probabilities of its productions.

The space of possible PCFGs for a given set of nodes is thus a multi-dimensional simplex; each non-terminal node $N_j$ has $K$ possible productions $N_j \rightarrow R_k$, each of which has a probability $0 \leq p_{jk} \leq 1$. With the constraint that $\sum_k p_{jk} = 1$, the space of possible vectors $(p_{j1}, \ldots, p_{jK})$ covers the $K$-dimensional unit simplex $\delta_K$. Since each non-terminal $N_j$ has some number $K(j)$ possible productions, the space of possible PCFGs is the Cartesian product of the simplices $(\delta_{K(1)}, \ldots, \delta_{K(J)})$. Note that this is a continuous, uncountable space, which makes inference in this space more complicated than in a discrete, finite space (see Section 4.4).

Also, note that a Dirichlet prior can be defined for a PCFG by using one Dirichlet prior for each non-terminal node. The Dirichlet prior defines a probability distribution over a simplex, so assigning one Dirichlet for each non-terminal results in a probability distribution over the whole PCFG.

$G$ consists of three types of productions:

1. Structural rules (e.g., $NP \rightarrow N$)
2. Semantic rules (e.g., $N \rightarrow \text{ball}$)
3. Lexical rules (e.g., $\text{ball} \rightarrow \text{ball}$)

The structural rules are the standard syntactic productions of a PCFG, taking one syntactic category and yielding one or more syntactic categories. These create the syntactic structure of the sentence. The semantic productions take in a syntactic category and return a referent, either an object or an action being depicted in the picture. Each of these productions must re-write a simple syntactic category (not a phrase) to a single referent. The lexical productions take in a referent and return a single lexical item, a terminal node in this PCFG. This results in trees like those depicted in Figure 2, where the terminal syntactic nodes immediately dominate semantic nodes, which in turn immediately dominate lexical nodes. The non-terminals $\text{cat}$, $\text{dog}$, and $\text{run}$ are the referents, while the terminals “cat”, “dog,” and “run” are the lexical items.
4.2.2 Trees and Semantics

Each tree $T_i$ thus contains all the syntactic, semantic, and lexical information for its sentence $S_i$. The sentence that a tree generates can be determined by reading off the terminal nodes of the tree in linear order. The proposition represented by the sentence is determined with only slightly more effort. The PCFG $G$ marks each NP as the subject or object of the sentence, and each NP contains a single noun-referent. The noun-referent contained in the Subject NP is considered the subject of the proposition, and the noun-referent contained in the Object NP is considered the object. Each sentence contains a single VP, and the verb-referent contained in it is the verb of the proposition. This means that each tree has a single unambiguous meaning in this grammar.

4.2.3 Generative Process

The model assumes that the input is generated as follows. First, the PCFG $G$ is drawn from a Dirichlet prior $\alpha$ that favors a sparse grammar, one with few of the productions having significant probability. A sparser grammar is a simpler grammar, with fewer rules for the learner to remember, and thus should be favored. Next, a picture $\pi$ is observed, and based on the propositions depicted in $\pi$ and the grammar $G$, a tree $T$ is drawn. The tree is drawn from the set of trees consistent with the picture $\pi$, based on the tree’s probability in the PCFG $G$. Thus a tree that is more probable in $G$ is more likely to be used to describe the picture $\pi$, assuming that the tree is consistent with $\pi$. Lastly, the sentence $S$ is read off of the terminal nodes of the tree $T$. This yields the graphical model shown in Figure 3.

Note that the acquisition model does not receive all of the data from this generative process. It can see the pictures $\pi$ and the sentences $S$, but it is not given the trees that generated the sentences, nor the grammar that generated the trees.\(^5\) Instead, it must infer a reasonable choice for the grammar from the training data it has observed and the generative process it assumes. But what does it mean to say that the acquisition model assumes a generative process? Simply put, it means that the acquisition model is trying to learn a grammar that makes sense for this generative process. It will prefer a grammar that is likely to have generated the observed data over one that is unlikely to have generated it. For instance, if the observed data is a large collection of sentences with Subject-Object-Verb word order, the model will favor a grammar that assigns high probability to this word order over a grammar that assigns high probability to Verb-Subject-Object word order.

The learning process uses Bayesian reasoning to infer probability distributions over the trees and the grammar. This is done with an alternating estimation scheme. Starting from an initial grammar $G_0$, the model chooses a tree-structure for each picture-sentence pair, based on the probabilities from $G_0$. These trees are then used in a Markov Chain Monte Carlo (MCMC) sampling scheme to propose a new grammar $G^*$, which will be accepted with some probability. If the proposed grammar is accepted, it becomes the new grammar $G_1$. If it is rejected, then the new grammar $G_1$ stays the same as the old one. New tree-structures are then chosen based on $G_1$, which in turn are used to propose a new grammar, and so on. The MCMC scheme is mathematically guaranteed to converge to a stable distribution, given enough time, so after a few hundred iterations, we can look at the grammars being generated from this scheme as approximating the true probability distribution over grammars.

4.3 Model plausibility

Before getting into the mathematical details of this implementation, we first consider the plausibility of the model as a model of child language acquisition. As with any model, a number of simplifications have had to be made in order to make the model tractable and inter-

\(^5\)The prior $\alpha$ is a special case. It is not observed, but it is not learned, either. Instead, it is set heuristically in the model at the outset.
pretatable, but we argue that these simplifications do not significantly impede the plausibility of the model — especially as the model is intended to be a computational level explanation. In this section, we discuss potential sources of implausibility and how they have been accounted for.

**Unsupervised Learning** Models can be either supervised or unsupervised. The model in this paper, like most other models of acquisition, is unsupervised, meaning that the model has no independent knowledge of the “right” answer. This is especially important with ambiguous pictures, where the model has no information about which proposition the sentence refers to, other than that it is some proposition in the current picture. Unsupervised models are more plausible at explaining human data, since, as discussed in Section 2.1.3, infants do not seem to get clear positive or negative evidence of grammaticality.

**Batch v. Online Learning** Models can look through data in a batch or online method. In a batch method, the model is presented with all of its training data at once, and it may recall any training example as many times as needed. In an online method, the model is exposed to each training example, one at a time, with no ability to look back or store the whole dataset (although some subset of the data may be stored). Clearly, online methods are a more accurate representation of the human acquisition problem. The current model uses a batch method to make its design and sampling methods clearer. However, the model could be converted to operate in an online setup by using a different sampling method, such as a particle filter (Doucet, Andrieu, & Godsill, 2000; Levy, Reali, & Griffiths, 2009), improving its cognitive plausibility.

**Memory Constraints** Another algorithmic concern with models is the amount of memory they require, as large memory requirements are generally assumed to be cognitively implausible (Goldwater, 2007, pg. 20). Gold-water notes that Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis method used in this model, are better than many other standard computational methods, such as Expectation-Maximization (EM), in terms of memory requirements. In addition, people learning to categorize objects can behave in a manner consistent with MCMC sampling (Sanborn & Griffiths, 2008), suggesting that even if MCMC itself is not cognitively plausible, humans are able to approximate it.

**Knowledge of Categories** The model makes the assumption that the learner already has available some notion of syntactic categories, since the PCFG relies on Subject, Verb, and Object categories. This may initially seem like the imposition of a language-specific structure; presumably, the ideal model would be able to learn these categories, rather than having them pre-specified. It is clearly difficult to determine that a child recognizes syntactic categories, but there is evidence that children do classify words into syntactic categories by 24–30 months (Valian, 1986), and proposed grammars of very early child language posit simple syntactic categories (e.g., Braine, 1963). Additionally, Spelke and Kinzler (2007) suggest that notions of agency and objecthood exist in infancy, which fits with our simple notions of Subject, Verb, and Object in our model. Thus, by the time that children are learning words in earnest, it seems reasonable to assume that they have some cognitive approximation of adult syntactic categories.

**Hierarchical structure** A related concern is that the PCFG automatically creates a hierarchical tree-structure for the sentence. Again, it may be that the ideal model would learn for itself that hierarchical structure should exist, rather than having it specified from the start. However, experimental evidence shows that by 18 months, children may be able to use adult-like syntactic structures to comprehend the meanings of anaphora (Lidz, Waxman, & Freedman, 2003), and by 3 years are using syntactic structure to determine permissible referents for pronouns (Crain & McKee, 1985). As with syntactic categories, it would appear that by the point in development that our model is considering, children already have some notion of hierarchical structure. Further, it should be noted that the PCFG used in this model has minimal hierarchical structure compared to most theories of adult grammars, and the model barely uses the hierarchical nature of the structure.

**Sentence simplification** The sentences used to train the model contain only content words; each word in the sentence maps to an object or action in the picture. This may seem like a simplification of the learning problem, since a child must learn which words are function words and which are content words. However, Jusczyk (1997) and Shafer, Shucard, Shucard, and Gerken (1998) have shown that children can differentiate function from content words at 10–12 months, based on frequency and prosody without necessarily knowing the words’ meanings. Thus, when the child is attempting to learn the meanings of words, the function words will presumably be filtered out, leaving sentences in the form that our model receives them. In fact, it is possible that excluding function words actually disadvantages the model, since it cannot take advantage of distributational cues that function words may offer.

**Referential uncertainty and noise** The model is designed so that one or more propositions may be true in the picture for a sentence. This is an improvement on the Mauritis et al. (2009) model, which assumed unambiguous contexts for its sentences. It does not span the full range of referential uncertainty, since the model assumes that each sentence describes some proposition visible in the presented picture. This, however, is standard both in other models of acquisition (Siskind, 1996; Piantadosi et al., 2008) and artificial language learning experiments.
This restriction could be relaxed by accounting for noise. The input a child receives is noisy in two ways: some of it is grammatically or factually incorrect, and some of it is irrelevant to the current state of the world. The model as presented here does not account for this noise directly. Although not presented here, a simple extension of the model in which some probability mass is assigned to false sentences would account for this noise and potentially improve the accuracy of the results. However, even without this extension, the present version of the model is sufficient for testing acquisition effects in artificial language learning experiments, where noise levels are generally quite low.

**Salience of References** The exhaustive list of all propositions depicted in a picture is extremely large, if not infinite – this is essentially a restatement of the gavagai problem. However, not all propositions are equally salient; social cues, foreground, novelty, and other effects cause certain meanings to stand out. The model cannot consider all the propositions in each picture, so instead it considers especially salient ones. A range of experiments, reviewed by Hollich et al. (2000), show that children display sensitivity to social and attentional cues during word-learning by 24 months old. This suggests that children exclude less salient propositions when early in acquisition, and are able to narrow down the range of possible propositions to a relatively small set of salient propositions. This salient set would then be similar to the input which our model receives.

### 4.4 Bayesian Inference

The core of the acquisition model is Bayesian inference, a probabilistic strategy to infer the rules underlying the language. (This discussion follows that of Griffiths & Yuille, 2006.)

In Bayesian inference, an agent is presented with some data $D$. This data can take a number of forms; in the context of language learning, it could range from full sentences for grammar induction to specific waveforms for phonetic learning. The agent wants to use this data to gain information about the underlying process that generated it. Let $H$ be a hypothesis about the generative process that led to $D$. Then the agent wants to learn the probability of this hypothesis given the data – $P(H|D)$. This probability, called the posterior, can be calculated using Bayes’ Rule:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{1}$$

The numerator consists of two terms: the likelihood, $P(D|H)$, and the prior, $P(H)$. The likelihood is defined as the probability of the observed data if the hypothesis $H$ is true. The prior represents the probability of $H$ being the correct explanation if the observed data were to be ignored. The denominator is the probability of the data, independent of the chosen hypothesis. This value can be calculated by marginalization over all hypotheses:

$$P(D) = \sum_{H'} P(D|H')P(H') \tag{2}$$

Note that $P(D)$ is independent of the specific hypothesis being considered, serving only as a normalization factor to ensure that the posterior probabilities are well-defined (i.e., that they sum to one). As such, one can think of Bayes’ Rule as stating that the posterior probability of a hypothesis is proportional to the likelihood times the prior.

Consider an extremely simple case of Bayesian reasoning involving a trick coin. The coin has the same image on both sides, either both heads (call this hypothesis $H_1$) or both tails (hypothesis $H_2$). Before you see the coin, you might suppose that each possibility is equally likely; thus your prior probability distribution is $p(H_1) = p(H_2) = .5$. You are then shown one side of the coin, and suppose that it is heads. This is the data $D$. The probability of seeing a head if both sides are heads ($P(D|H_1)$) is one, while the probability of seeing a head if both sides are tails ($P(D|H_2)$) is zero. So the probability that the coin is two-headed, given the observed data that it is at least one-headed is

$$P(H_1|D) = \frac{P(D|H_1)p(H_1)}{\sum_{H'} P(D|H')P(H')} = \frac{1 \cdot .5}{(1 \cdot .5) + (0 \cdot .5)} = 1,$$

because the only possibility is that the coin is two-headed. Now suppose that you have a standard coin as well. You take the trick and the standard coin, mix them up in your pocket, and pull one out at random. The prior probability that the coin you have pulled out is the trick coin is one-half, as is the prior that it is the standard coin. Call these hypotheses $H_T$ and $H_S$. Again, you see one side of the coin you have pulled out, and again it is heads. What is the probability that you have pulled out the trick coin? The probability of the observed data given the trick coin is one, since both sides are heads, and the probability of the data given the standard coin is one-half, since one side is heads and one is tails. The probability that you have drawn the two-headed coin is now

$$P(H_T|D) = \frac{P(D|H_T)p(H_T)}{\sum_{H'} P(D|H')P(H')} = \frac{1 \cdot .5}{(1 \cdot .5) + (.5 \cdot .5)} = \frac{2}{3},$$

and the probability that you have drawn the standard coin is
\[
P(H_S|D) = \frac{P(D|H_S)P(H_S)}{\sum_{H'} P(D|H')P(H')}
\]

\[
= \frac{.5 \cdot .5}{(1 \cdot .5) + (.5 \cdot .5)} = \frac{1}{3}.
\]

These situations have a finite number of hypotheses, but Bayes’ Rule can be extended to infinite hypothesis spaces as well by replacing the sum over hypotheses with an integral:

\[
P(H|D) = \int H' P(D|H')P(H')dH'
\]  \hspace{1cm} (3)

This form is used to infer grammars in the model, since each possible weighting of the PCFG is a different hypothesis, and there are uncountably infinitely many possibilities. As will become clear in the next section, the exact solution of the inference problem is intractable, and a Markov Chain Monte Carlo approximation method will be substituted.

4.5 The Learning Model

The model’s goal is to find reasonable values for the probabilities in the PCFG \( G \), based on the observed data and the imposed Dirichlet prior over \( \alpha \). To do this, the model alternates between using the current estimate of \( G \) to select trees \( T \) for the picture-sentence pairs, and using the trees \( T \) to infer a new estimate of \( G \). The model uses a Metropolis algorithm to produce PCFG weights in \( G \) and samples trees \( T \) from the estimated PCFG (staying consistent with \( \pi \) and \( S \)). To start the process, trees are chosen from a uniform PCFG. \( T \) and \( G \) are repeatedly chosen until their values begin to converge, usually after a few hundred iterations.

4.5.1 Sampling \( T \)

We start with the comparatively simple step of sampling the trees \( T = \{T_i\} \) using the PCFG \( G \). We sample using a product of experts function, multiplying the likelihood of the tree in the current grammar \( G \) by the consistency of the tree for the observed picture and sentence pair \( (\pi, S_i) \). This “consistency” measure is a characteristic function \( \chi(T_i; \pi, S) \), equal to 1 if the meaning of \( T_i \) is consistent with some proposition pictured in the picture \( \pi \), and if its terminal nodes match the sentence \( S_i \), and equal to zero otherwise.

\[
T_i \sim p(T_i|G, \pi, S_i)
\]

\[
\sim \frac{p(T_i|G)\chi(T_i; \pi, S)}{\sum_{T'} p(T'|G)\chi(T'; \pi, S)}
\]  \hspace{1cm} (4)

\( p(T_i|G) \) is calculated from the PCFG probabilities in \( G \), and the consistency function is easily calculated from the pictures and sentences. Note that, unfortunately, we still need to calculate the normalization term for the distribution that \( T_i \) is drawn from, because multiplying by the consistency function leads to a loss of probability mass from tree inconsistent with the picture or the sentence. This remains tractable as long as all the possible trees \( T' \) that are consistent with the picture and sentence can be enumerated.

4.5.2 The Probability of \( G \)

With the trees chosen, we infer the grammar. The first step is to calculate the probability of a grammar given the trees, pictures, and hyperparameters, using Bayes’ Rule.

\[
P(G|T, \alpha, \pi) = \frac{P(T|G, \alpha, \pi)P(G|\alpha)}{\int_G P(T|G, \alpha, \pi)P(G|\alpha)dG}
\]  \hspace{1cm} (5)

\( P(G|\alpha) \) is easy to calculate, since we have assumed that \( \alpha \) is a Dirichlet prior on the grammar. We assume a symmetric Dirichlet prior for each non-terminal. Recall that \( G \) is a set of probabilities \( \{g_{jk}\} \), where \( g_{jk} \) is the probability in the PCFG of the production \( N_j \rightarrow R_k \). The probability of the vector \( g_1 = (g_{j1}, \ldots, g_{jK}) \) ranging over all right-hand sides for the non-terminal \( N_j \), given the Dirichlet prior over \( \alpha \), is

\[
P(g_j | \alpha) = \prod_k g_{jk}^{\alpha-1}
\]

The probability of the whole PCFG is just the product of \( P(g_j | \alpha) \) over all left-hand sides \( N_j \):

\[
P(G | \alpha) = \prod_j g_{jk}^{\alpha-1}
\]  \hspace{1cm} (6)

This gives us half of Eqn. 5, the prior on \( G \). Calculating the likelihood term is a bit trickier. We start by looking at the probability of a single tree \( T_i \) given the grammar. As shown in the graphical model in Figure 3, \( T_i \) and \( \alpha \) are conditionally independent given the grammar, so we can remove \( \alpha \) and calculate the likelihood using the product of experts formula:

\[
P(T_i|G, \alpha, \pi_i) = \frac{P(T_i|G, \pi_i)}{\sum_{T'} P(T'|G)\chi(T'; \pi_i)}
\]

The sum in the denominator is over all possible trees, although only trees that are consistent with the picture \( \pi \), with contribute mass to the sum. Letting \( n_{jk}^{(i)} \) be the number of times that production \( N_j \rightarrow R_k \) is used in tree \( T_i \), and letting \( n_{jk}' \) be the same count for tree \( T' \), we can re-write the likelihood as

\[
P(T_i|G, \alpha, \pi_i) = \frac{\chi(T_i; \pi_i) \prod_{j,k} g_{jk}^{n_{jk}^{(i)}}}{\sum_{(T', \chi(T'; \pi_i)=1)} \prod_{j,k} g_{jk}^{n_{jk}^{(i)}}}.
\]  \hspace{1cm} (7)

The probability \( P(T|G, \pi) \) can be calculated by multiplying the probabilities of each tree \( T_i \); no additional
normalization is needed because the trees are conditionally independent given \( G \). Let \( N_{jk} = \sum n_{jk}^{(i)} \) be the total number of times each production occurs in the current treebank. Then the numerator of Eqn. 5 becomes

\[
P(T | G, \alpha, \pi) P(G | \alpha) = \prod_{j,k} g_{jk}^{\alpha_{jk} - 1} \prod_{i} \frac{\chi(T_i; \pi_i) \prod_{j,k} g_{jk}^{n_{jk}^{(i)}}}{\sum_{T'} \chi(T'_{i+1}; \pi_{i+1}) \prod_{j,k} g_{jk}^{n_{jk}^{(i)}}} = \prod_{i} \frac{\chi(T_i; \pi_i) \prod_{j,k} g_{jk}^{N_{jk}^{(i)} + \alpha_{jk} - 1}}{\sum_{T'} \chi(T'_{i+1}; \pi_{i+1}) \prod_{j,k} g_{jk}^{N_{jk}^{(i)}}}.
\]

(8)

The denominator of Eqn. 5, as the normalization factor, is just the integral of the above expression over all possible grammars \( G \):

\[
\int_G P(T | G, \alpha, \pi) P(G | \alpha) dG = \int_G \frac{\chi(T; \pi) \prod_j g_{jk}^{N_{jk}^{(i)} + \alpha_{jk} - 1}}{\sum_{T'} \chi(T'_{i+1}; \pi_{i+1}) \prod_{j,k} g_{jk}^{N_{jk}^{(i)}}} dG
\]

(9)

Note that the value of this integral is independent of the current value of the grammar, since it is an integral over all possible values of the grammar. This integral is intractable, and makes the equation for \( P(G | T, \alpha, \pi) \) unwieldy. However, due to the constant nature of the integral, the ratio \( P(G_1 | T, \alpha, \pi) / P(G_2 | T, \alpha, \pi) \) is tractable. Markov Chain Monte Carlo algorithms, specifically the Metropolis method, allow us to use this ratio to approximate the distribution we are interested in.

4.5.3 Choosing \( G \) with MCMC

In order to choose a grammar based on the trees, pictures, and hyperparameters, we need to find \( P(G | T, \alpha, \pi) \), but in the last section, we found this value to involve a complicated integral. However, there is a class of methods that can approximate pernicious probability functions: the Markov Chain Monte Carlo (MCMC) methods. MCMC methods can be used to generate samples from an intractable probability distribution and to estimate the expectation of functions under a given probability distribution (Mackay, 2005, pg. 357). We use an MCMC method in its first role in this section, and will use it in its second role in Section 4.6.

We use the Metropolis-Hastings algorithm to generate our samples from \( P(G | T, \alpha, \pi) \). Metropolis-Hastings (or Metropolis in the symmetric case we will consider here) is useful because it uses the unnormalized probability, which was obtained in Eqn. 8 and is tractable, instead of the normalized probability. Metropolis sampling consists of proposing a new grammar, based on the current grammar, and probabilistically accepting the new grammar depending on how its likelihood compares to the likelihood of the current grammar. Proposing a new grammar is also done probabilistically, using a proposal function \( Q \), where \( Q(G_2 | G_1) \) is the probability of proposing \( G_2 \) if the current grammar is \( G_1 \). Let \( G^{(t)} \) be the current grammar, and let \( G' \) be the newly proposed grammar. The probability of accepting \( G' \) is given by

\[
a = \frac{P(G' | T, \alpha, \pi)}{P(G^{(t)} | T, \alpha, \pi)} \cdot \frac{Q(G^{(t)} | G')}{Q(G' | G^{(t)})},
\]

with acceptance certain if \( a \geq 1 \). If the proposed grammar is accepted, set the new grammar \( G^{(t+1)} = G' \). If the proposed grammar is rejected, set the new grammar \( G^{(t+1)} = G^{(t)} \). Using a symmetric proposal distribution — one that is as likely to propose grammar \( G_1 \) from grammar \( G_2 \) as to propose grammar \( G_2 \) from grammar \( G_1 \) — simplifies matters because the second fraction equals 1 (as \( Q(G^{(t)} | G') = Q(G' | G^{(t)}) \)).

Note that these samples form the grammar are dependent, since the new grammar depends directly on its predecessor. By looking only at every \( n \)-th Metropolis sample, for sufficiently large \( n \), the samples become essentially independent. This allows us to generate samples from the intractable distribution \( P(G | T, \alpha, \pi) \), because the acceptance function does not require the calculation of the intractable integral in Eqn. 9. If we let \( G = \{ g_{jk} \} \) be the current value of the grammar, and \( G' = \{ g'_{jk} \} \) be the proposed grammar, and assuming we have a symmetric proposal distribution, then the acceptance function is

\[
P(G' | T, \alpha, \pi) = \frac{\chi(T; \pi) \prod_{j,k} g_{jk}^{N_{jk}^{(i)} + \alpha_{jk} - 1}}{\sum_{T'} \chi(T'_{i+1}; \pi_{i+1}) \prod_{j,k} g_{jk}^{N_{jk}^{(i)}}} = \frac{\prod_{j,k} g_{jk}^{N_{jk}^{(i)} + \alpha_{jk} - 1}}{\sum_{T'} \chi(T'_{i+1}; \pi_{i+1}) \prod_{j,k} g_{jk}^{N_{jk}^{(i)}}}.
\]

(10)

This form for the acceptance function is contingent on the identification of a symmetric proposal distribution; with an asymmetric proposal distribution, a corrective factor would need to be added. Luckily, a symmetric proposal distribution is available.

4.5.4 Softmax Symmetric Distribution

The space of possible PCFGs is complicated because it is a Cartesian product of simplices. Simplices are bounded, so many common symmetric proposal distributions will not work, as they will propose points outside of the simplex. We can get around this issue by moving to softmax space. Softmax space is a copy of \( \mathbb{R}^K \) (which is unbounded) with a mapping to the \( K \)-dimensional simplex. The mapping is defined as follows; for each vector
Because the softmax space is unbounded, Gaussian distributions can be used there. Gaussians are symmetric, so the Metropolis acceptance function (Eqn. 11) holds in softmax space. This allows us to run the Metropolis algorithm in softmax space. Since we can move from softmax space to the space of possible PCFGs trivially using Eqn. 12, we can run the Metropolis algorithm and extract the grammar from softmax space whenever it is needed (e.g., in estimating the grammar from softmax space whenever it is needed).

However, to do this, we need to work out the softmax acceptance function. Conveniently, softmax acceptance function and the PCFG acceptance function are the same. To see this, let \( H \) be the current softmax grammar, and let \( H' \) be the proposed softmax grammar. These softmax grammars generate the PCFGs \( G \) and \( G' \). We want to show that

\[
\frac{P(H'|T,\alpha,\pi)}{P(H|T,\alpha,\pi)} = \frac{P(G'|T,\alpha,\pi)}{P(G|T,\alpha,\pi)} \frac{Q(G;G')}{Q(G;G)}
\]

Let \( m : H \mapsto G \) be the mapping from softmax space to PCFG space given by Eqn. 12. Changing the variable in a probability distribution requires a corrective term to keep the probability well-defined (i.e., with integral 1); if there is a bijective function \( h \) such that \( y = h(x) \), and a probability distribution \( f \) over \( y \), then the equivalent probability function \( g \) over \( x \) is given by

\[
g(x) = f(y) |J(h)(x)|,
\]

where \( J(h)(x) \) is the Jacobian determinant of \( h \), evaluated at \( x \). We want to change the variable \( H \) to \( G \), using the mapping \( m \), so

\[
P(G|H|T,\alpha,\pi) = P(m(H)|T,\alpha,\pi) = P(H|T,\alpha,\pi) |J(m)(H)|.
\]

And, since \( Q(H';H) \) is defined to be a Gaussian centered at \( H \), we have

\[
Gauss_H(H') = Gauss_m(m(H')) = Gauss_{G'}[J(m)(G')].
\]

Note that \( m(H) \) does not require a corrective term because it is the mean of the Gaussian.

These facts combine to show that

\[
\frac{P(H'|T,\alpha,\pi)}{P(H|T,\alpha,\pi)} \frac{Q(H;H')}{Q(H';H)} = \frac{P(G'|T,\alpha,\pi)}{P(G|T,\alpha,\pi)} \frac{|J(m)(H')|}{|J(m)(H)|} \frac{Q(G;G')}{Q(G';G)} \frac{J(m)(H)}{J(m)(H')}
\]

The last point is that \( m \) is not bijective over the whole softmax space. If \( c \) is a constant term added to all dimensions of the softmax space, then \( m(H) = m(H + c) \), because

\[
\frac{\sum_{k'} e^{h_{jk}} e^{h_{jk} + c}}{\sum_{k'} e^{h_{jk}} e^{h_{jk} + c}} = \frac{\sum_{k'} e^{h_{jk}}}{\sum_{k'} e^{h_{jk}} e^{c/k'}}
\]

To create a bijective mapping, we constrain \( H \) by requiring that \( \sum_k h_{jk} = 0 \). This can be done by drawing the \( h_{jk} \) from independent Gaussians and subtracting the mean over \( h_j \). This subtraction retains symmetry and restricts the domain so as to render \( m \) bijective.

### 4.5.5 The Complete Learner

The main result here is that we can perform Metropolis sampling in softmax space by drawing a proposal softmax grammar \( H' \), converting it to a PCFG \( G' \), and accepting the proposed grammar with probability

\[
\left( \prod_{j,k} \frac{P(H'|T,\alpha,\pi)}{P(H|T,\alpha,\pi)} \right) \frac{Q(G';G)}{Q(G;G')} \frac{J(m)(H)}{J(m)(H')}
\]

The new grammar is then used to select a new set of trees using

\[
T_i \sim \frac{p(T_i|G)(T_i;\pi,S_i)}{\sum_{T'} p(T'|G)(T'|\pi,S_i)}
\]

This is repeated until the grammar settles, which usually occurs within 1000–2000 iterations. The Metropolis algorithm converges to a stationary distribution, rather than a single element, so there is no “correct” grammar. Instead, the final output of the model is a distribution over likely grammars. In the cases investigated in this paper, the distribution is relatively peaked — a small subspace of the possible grammars are significantly better than the rest. These grammars are those closely match the grammar that generated the input. As such, when reporting results, we can generally look at a single grammar, or the mean of a few grammars, as a representative grammar for the learned distribution.

### 4.6 Online Expectation Calculation

The model can be used to investigate more than just the acquisition process. In addition, we can look at how the grammar that is learned affects behavior in linguistic tasks. One linguistic task that readily admits to analysis under this model is the ability to predict the next word from words already seen in a sentence. This is a core component of sentence processing and parsing, and is a cornerstone of psycholinguistic research.

In this situation, we observe a sentence prefix, the first \( i \) words of a sentence, \( w_1^i = w_1, w_2, \ldots, w_i \) and want
to make predictions about likely continuations. This behavior is common in adults and is a cornerstone of psycholinguistic research. We want to investigate the predictions made by our model in this task. Thus we want to calculate \( p(w_{i+1}|w_i, \text{training}) \).\(^6\) We start by marginalizing over the specific grammar \( G \):

\[
p(w_{i+1} = t|w_i) = \int_G p(w_{i+1} = t|w_i, G)p(G)dG
\]

This assumes that the learner is entertaining not a single grammar, but rather a distribution over possible grammars. This integral is intractable, but Markov Chain Monte Carlo methods, such as Metropolis sampling, yield approximations of difficult integrals. Specifically, an integral of the form \( \int f(x)p(x)dx \), can be approximated by a MCMC method that draws samples of \( x \) from the distribution \( p(x) \). As the number of drawn samples \( \{x_1, \ldots, x_n\} \) grows, the average value of \( f(x_i) \) over all of the samples tends to the value of the integral:

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f(x_i) = \int f(x)p(x)dx
\]

Thus, if we draw \( N \) samples \( \{G_1, \ldots, G_N\} \) from the Metropolis sampler, we have

\[
\int_G p(w_{i+1} = t|w_i)p(G)dG \approx \frac{1}{N} \sum_{j=1}^{N} p(w_{i+1} = t|w_i, G_j)
\]

The next step is to break down the probability \( p(w_{i+1} = t|w_i, G_j) \):

\[
\frac{1}{N} \sum_{j=1}^{N} p(w_{i+1} = t|w_i, G_j)
\]

\[
= \frac{1}{N} \sum_{j=1}^{N} \frac{p(w_{i+1}, w_{i+1} = t|G_j)}{p(w_i|G_j)}
\]

\[
= \frac{1}{N} \sum_{j=1}^{N} \sum_{t'} p(w_{i+1} = t|G_j)
\]

\[
= \frac{1}{N} \sum_{j=1}^{N} \sum_{t'} p(w_{i+1} = t|G_j)
\]

The sum in the denominator is over all words \( t' \) that could follow the prefix. Next, we marginalize again; this time the marginalization is over propositions \( r \) that the sentence may be expressing. The probability \( p(r) \) is learned from the world-knowledge, and in the current model is a smoothed multinomial model whose parameters are set by observation. The last marginalization is over the possible structural realizations \( s \) of the proposition \( r \); each proposition can occur in SVO order, SOV order, and so on:

\[
\frac{1}{N} \sum_{j=1}^{N} \sum_{t'} \sum_{s} p(w_{i+1} = t|G_j)
\]

The probability of a sentence string given the structural realization of its proposition is a straightforward calculation; it is the product of each of the semantic items going to the appropriate word, based on their order. For instance, if the proposition is \[ \text{Subj} = \text{MOM}, \text{Verb} = \text{EAT}, \text{Obj} = \text{CEREAL} \], and the structural realization has (incorrect) SOV order, then the likelihood of “Mom eats cereal” given this structural realization is

\[
p(\text{MOM} \rightarrow \text{Mom})(\text{CEREAL} \rightarrow \text{eats})(\text{EAT} \rightarrow \text{cereal}),
\]

whereas the likelihood of “Mom eats cereal” under the correct SVO realization is

\[
p(\text{MOM} \rightarrow \text{mom})(\text{EAT} \rightarrow \text{eats})(\text{CEREAL} \rightarrow \text{cereal}),
\]

This gives us our final form for the expectation probabilities:

\[
p(w_{i+1} = t|w_i) =
\]

\[
= \frac{1}{N} \sum_{j=1}^{N} \sum_{t'} \sum_{s} p(s, t, r, s, G_j)p(s|G_j)p(r|s|t)
\]

\[
= \frac{1}{N} \sum_{j=1}^{N} \sum_{t} \sum_{s} p(s, t, r, s, G_j)p(s|G_j)p(r|s|t)
\]

Note that the calculation of these probabilities separates the linguistic knowledge (i.e., the grammar) from the world knowledge (i.e., the probability of each proposition). This separation can be of varying degrees. In the experiments in this paper, the separation is nearly complete, as the grammar is unlexicalized and both the subject and object first re-write as a noun before re-writing semantically (see the next section for details). This presents a strong hypothesis that world knowledge affects the grammar only indirectly. At the other extreme, the model could learn a strongly lexicalized PCFG\(^7\), in which case world knowledge would be a core component of the grammar. The middle ground between these two positions is probably best, and we will return to this point in Section 6.

With this equation and those of the preceding section, we can investigate both the grammar that the model learns and investigate some ways that these grammars affect sentence processing and other linguistic tasks.

\(^6\)“Training” here refers to the accumulated knowledge about the grammar, lexicon, and the world. We will omit the phrase in the equations.

\(^7\)A lexicalized PCFG is one where the non-terminal nodes are labelled not only with their syntactic categories, but also their lexical heads (e.g., Charniak, 1997; Collins, 1999).
5 Experiments

We perform four experiments to investigate the four purposes of the model listed in Section 3. The first experiment looks at the model’s robustness to ambiguity. The second shows that the model is able to store world knowledge about likely agents and patients outside of the PCFG. The last two experiments look at the model’s ability to explain the results from Amato and MacDonald (2008); one uses the same setup as in their experiments, while the other looks at how the model’s behavior changes with extra non-linguistic data.

Shared setup All of the experiments look at simple transitive sentences, and all of the experiments use the same PCFG structure. The root node is ternary branching, with children NP-SUBJ, VP, and NP-OBJ for subject, verb, and object, respectively. The object is separate from the VP to allow for VSO and OSV word orders. Each of these phrases rewrites to a terminal category — VP → V, NP → N — which then rewrites to the semantic node. Thus the semantic nodes and subject/object roles are independent, and the PCFG will not contain any information about the likelihood of a certain referent taking a certain semantic role (subject or object). This is essential for testing the separation of linguistic and world knowledge. Additionally, no words or meanings are repeated in a single sentence; reflexive sentences like baby see baby are not allowed.

In each of the experiments, we run through 2000 iterations of the Metropolis algorithm to reach a stable distribution over grammars. Samples of the grammar are drawn every 10 iterations in order to approach independence. Because the search space of possible grammars is quite large, the model will often become stuck in local optima. As such, each run of the model starts with five 2000-iteration Metropolis chains. The best\textsuperscript{8} chain is kept and the others discarded. The Metropolis algorithm uses a Gaussian in softmax space to draw its proposed grammars, which introduces a single parameter into the model: the Gaussian’s variance, $\sigma^2$. Changing this variance affects the rate of convergence, with smaller variances slowing convergence but decreasing the likelihood of becoming stuck in substantially sub-optimal regions of the search space. We use $\sigma^2 = 0.1$ in these experiments, a value chosen heuristically for effectively balancing rate and quality of convergence. Additionally, the Dirichlet hyperparameter $\alpha$ is set to 1, neither favoring nor disfavoring sparsity. Other values affected the sparseness of the grammar but did not noticeably affect the accuracy of the learned grammar.

\textsuperscript{8}Best here refers to the grammar whose 2000th iteration assigns the greatest likelihood to the observed sentences, conditioned on the observed pictures and marginalized over all trees.

<table>
<thead>
<tr>
<th>Lexical</th>
<th>Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.2%</td>
<td>90%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word Order Probs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVO</td>
</tr>
<tr>
<td>84.5 ± 6.0%</td>
</tr>
</tbody>
</table>

5.1 Robustness to Ambiguity

The first experiment addresses the model’s ability to learn linguistic structure from ambiguous data, with ambiguity both in the meaning (multiple propositions per picture) and the sentence structure (VSO and SOV word order alternation). The input consists of 35 randomly generated transitive picture-sentence pairs. The lexicon for this input has seven nouns and seven verbs. Each picture depicts three propositions, one the intended meaning of the sentence and the other two randomly chosen. This induces referential uncertainty. To create syntactic uncertainty, there is a 75\% chance that a sentence will have VSO word order and a 25\% chance it will have SOV word order; the artificial language is thus similar to a language like German, which uses SOV word order in subordinate clauses and SVO otherwise. A word is considered to be learned correctly if it is the most likely rewrite for its referent, so the word cat has been learned correctly if it is the most likely lexical item to describe the semantic meaning cat\textsuperscript{9}.

Table 1 shows the success rates of the model, averaged over 50 trials. Lexical accuracy looks at the by-word learning rates; of the 700 words to be learned over the 50 trials, only 5 were not successfully learned. A more useful measure of the learning ability of the model is the overall lexicon accuracy. Lexicon accuracy counts only models where all 14 words were correctly learned; 45 of the 50 trials had a completely accurate lexicon, and each of the trials with an incorrect lexicon was off by a single word.

Furthermore, the model’s estimate of the probabilities of the two word orders are qualitatively correct; SVO is favored over SOV, and none of the other word orders have substantial probability mass assigned to them. The probabilities do not quantitatively match the input distribution, but are close to it.

5.2 World Knowledge

The second experiment looks at the encoding of world knowledge in the model. We investigate the effect of having certain referents being good subjects or good objects. Such a situation can arise in the real world due to animacy differences; humans and other animals are common
Table 2: Disambiguating subjects and objects.

<table>
<thead>
<tr>
<th>Subject Noun</th>
<th>Verb</th>
<th>Object Noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2 ± 1.0%</td>
<td>8.5 ± 1.5%</td>
<td>84.2 ± 10.4%</td>
</tr>
</tbody>
</table>

Table 3: Online expectations generated from Amato and MacDonald’s input data, based on 150 trials.

<table>
<thead>
<tr>
<th>Preferred Object</th>
<th>Other Objects</th>
<th>Other Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.1 ± 11.3%</td>
<td>20.1 ± 7.3%</td>
<td>2.6 ± 1.0%</td>
</tr>
</tbody>
</table>

agents, while buildings, furniture, and other inanimate objects are common patients. This sort of information is used in sentence processing; Trueswell et al. (1994) showed that readers were slower in reading reduced relative clauses with an animate NP (the defendant examined by the lawyer) than an inanimate NP (the evidence examined by the lawyer), because of the initially plausible analysis that the animate NP was the subject in a main verb reading of the sentence (e.g., the defendant examined the evidence). We investigate the ability of our model to pick up on this information and use it in expectation generation.

The training data for this experiment contains six nouns and three verbs. The input consists of 18 sentence-picture pairs, each with an unambiguous meaning. Three of the nouns only occur as subjects in the training pictures, and the other three nouns appear only as objects. The sentences all use SVO word order. Recall that NP-SUBJ or NP-OBJ re-writes first as N, and N re-writes as the semantic node. This means that noun-specific subject/object biases cannot be encoded in the grammar. Instead, this bias arises from the world knowledge.

Using the online expectation formula from Section 4.6 shows that the model is able to differentiate common objects from subjects. Table 2 shows the likelihood of each class of words completing a sentence that begins with a subject noun followed by a verb. Because this is a strictly SVO language, we expect to see a strong preference for an object noun to complete the sentence. The probability of each class is the sum of the average probability of each word in that class over 150 trials. Note that subject nouns and verbs are equally unlikely in this model, presumably as neither has occurred as an object. A better model of proposition likelihood may remedy this false equivalence; presumably subject nouns should still be somewhat more likely than verbs.

5.3 Second-order Dependencies

Building off the online expectation results above, we look at the ability of learners to use more complicated world knowledge in sentence processing; second-order dependencies. Amato and MacDonald (2008) tested adults learning a simple artificial language where certain subject-verb-object combinations were more common than others. Their language contained three possible subjects, three possible verbs, and three possible objects. Each subject-verb pair occurred nine times in their training set, seven times with one of the objects, and once each with the other two. The preferred object varied depending on the subject-verb choice, so each object was preferred by three subject-verb pairs. Training in the artificial language consisted of the display of a picture showing a single proposition (one subject-verb-object action) and a sentence describing the picture using the artificial language. This is illustrated in Figure 1; the two bottom lines of text, showing the English translation, are not shown to the learner. Note that the sentences in the artificial language are consistently verb-subject-object.

After training, the learners participate in a self-paced reading time task in the artificial language. This task consists of reading sentences in the absence of pictures, so the learners cannot directly see what proposition the sentence is describing. Because the sentences have VSO order, the learners know the subject-verb pair before reading the object, and the experiment showed that reading the preferred object for a subject-verb pair is done significantly faster than reading a dispreferred object. Thus the learners are able to quickly use world knowledge about likely states of the world in sentence processing. We show that our model displays the same effect.

We recreate Amato and MacDonald’s corpus in this experiment, with three subject nouns, three verbs, and three object nouns. The preferred object for each subject-verb pair is shown seven times, and the dispreferred object is shown only once. Following Amato and MacDonald, the pictures are unambiguous. Table 3 shows the online expectations generated by the model; note that the model finds the preferred object to be a more likely continuation than other objects. In fact, the probability mass assigned to the dispreferred objects is not significantly different from the probability mass assigned to another subject or verb completing the sentence, even though the dispreferred objects have been seen to occur with the subject-verb prefix. This may be a result of uncertainty about word meanings; the non-object words may gain probability mass from being confused with object meanings.

5.4 Learning from Non-Linguistic Data

The last experiment uses the same setup as the previous experiment, but looks at learning directly from world knowledge without linguistic information. Again, we are interested in the ability of the learner to use second-order dependencies in sentence processing. In this experiment, though, we induce the object preferences strictly through world knowledge.

The input for this experiment is the same as in the preceding experiment, but each sentence is seen only once.
Table 4: Online expectations generated from Amato and MacDonald’s input when world knowledge alone induces object bias.

<table>
<thead>
<tr>
<th>Preferred Object</th>
<th>Other Object</th>
<th>Other Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.3 ± 11.5%</td>
<td>13.1 ± 9.7%</td>
<td>5.6 ± 10.1%</td>
</tr>
</tbody>
</table>

To induce the object preferences, the model observes an additional six pictures showing each of the preferred propositions. Thus the model sees the same pictures as in the previous experiment, showing a preference for certain objects given a subject-verb pair, but the linguistic input is unbiased. Again, the results are calculated over 150 trials.

Table 4 shows that the model still learns that the preferred objects are more likely completions than the dispreferred objects or other words. This shows that, within the model, world knowledge gained from non-linguistic experience can be used in linguistic judgments, although the effect is not as strong as when this non-linguistic knowledge is also used in creating the grammar.

6 Discussion

The model proposed in this paper successfully learns the syntax and lexicon of a language from ambiguous data. It offers a paradigm for storing world knowledge and allowing that world knowledge to be used in linguistic judgments in ways that are at least qualitatively the same as human language learners. The model thus gives us a computational-level framework for learning and using a grammar of a new language. In this section, we consider what this framework reveals and discuss some questions it raises.

6.1 The Model and the Learning Problem

First, we must consider what the model does and does not reveal about the learning problem. The model presents a computational-level solution to the problem of learning a language from context-dependent utterances. It shows that statistical properties can be used to infer both the syntax and lexicon of a language simultaneously. Returning to the issues in language learning addressed in Section 2.1.3, we see that the model presents further evidence of their surmountability. The model’s ability to infer the correct meanings for the lexicon without any unambiguous input gives further credence to the argument that cross-situational learning can be used to overcome the gavagai problem. Also, note that the model never encounters any direct negative evidence; all that it knows is whether the tree-structure is given to a sentence is valid for the observed picture, and even this validity is checked against its internal grammar and lexicon. There is no external source to correct the model; this is perhaps a stronger condition than that on children learning a language, as children are at least sometimes corrected on their grammar. The lack of negative evidence is overcome through the size principle; sparser grammars are favored because they do not waste probability mass on unobserved sentences and trees. Lastly, the joint acquisition task of the model uses implicit bootstrapping; each bit of information gained about the syntax constrains the lexicon, and vice versa. Thus the model uses bootstrapping from the start of the acquisition problem, without any explicit specification of a bootstrapping mechanism. This circumvents the chicken-or-egg nature of more formal semantic or syntactic bootstrapping mechanisms.

Furthermore, the model offers a new way of incorporating world knowledge into the grammar of a language. The world knowledge can be stored outside of the syntactic knowledge, which is useful in capturing second-order dependencies, which are difficult to store in the syntax even with subcategorization frames or lexicalized PCFGs. The model makes predictions about expectations in sentence processing that agree with observations of actual human sentence processing times. The model’s framework may also prove useful in computerized language learning. Rather than attempting to store more world knowledge in a lexicalized PCFG, this model offers a second path, where world knowledge is stored outside of the grammar and incorporated during parsing.

That said, the model does come with some caveats. The most prominent is that it is a computational-level solution. It shows that human language learners can use statistical methods to overcome ambiguity in their acquisition of a language. However, as a computational-level solution, it does not address the specific mechanisms that humans would use to implement this solution. The current implementation of the model uses a Markov Chain Monte Carlo method that, while cognitively plausible, likely differs from the learning method used by the brain. As a result, it is difficult to get quantitative predictions from this model that can be directly compared to human language usage. Instead, we are interested in more qualitative comparisons, like those of Section 5.3, that suggest the model is a step forward in uncovering the mechanisms of language acquisition. These qualitative predictions are eminently testable and, especially with future work, will provide significant evidence for or against the plausibility of the model.

We now consider a few outstanding questions raised by the model.

6.2 How Separate is World Knowledge?

The model shows that some world knowledge can be stored separately from the linguistic knowledge, but what should be in the grammar, and what should be separate? Using word-specific syntactic information has a long history in linguistics, whether in a semi-syntactic form like a subcategorization frame or in lexicalized grammars like HPSG. It is clear that word choice can influence syntactic
6.3 The Strength of Non-Linguistic Learning

The separation of world and linguistic knowledge introduces a new question: should they have equal input in sentence processing, or does one outweigh the other? As shown in the last experiment, the model learns from world knowledge in the absence of linguistic input, but should world knowledge be discounted somehow if it is not learned in the presence of language? As it stands, world knowledge gained in the presence of language and without its presence are weighted equally. But the presence or absence of linguistic accompaniment is probably a useful signal for sentence processing. Suppose there is something that often happens in the world, but is rarely remarked upon — blinking is one such occurrence. We have extensive world knowledge saying that everyone blinks, and that they do so quite often. Thus, in the model as currently implemented, a proposition like “The man blinks” is heavily weighted, even though an actual statement to that effect is rarely uttered. This is partially accounted for by the notion of salience; blinking is not a salient activity and thus is unlikely to merit a mention in the conversation. It may be useful to learn that situations that go unremarked are less salient.

One possible implementation of this is to weight propositions differently depending on whether there is linguistic data available. When there is no utterance to accompany a picture, all of the propositions are given additional probability mass. When there is an utterance, the proposition described by the utterance is given additional probability mass, whereas those propositions depicted in the picture but not described by the utterance lose probability mass. In this design, propositions that are likely to be true in the world gain probability mass overall, but only the ones that are true and often mentioned gain substantial probability mass.

6.4 Getting to 100%

One other issue is that our model does not generally learn a perfectly accurate lexicon; in most runs, at least one word was learned incorrectly. This is commonplace in models, both of word learning (Maurits et al., 2009) and other linguistic tasks (e.g., Goldwater, 2007). Yet human language learners manage to learn words quickly and very accurately. Can the model reach such high accuracy rates, and if so, how?

One possibility is that the model is limited by its implementation. Because of the complicated integrals and probability distributions in this and other models, we have to resort to MCMC sampling. It is conceivable that the brain would use a better estimation scheme, or be able to run it longer, avoiding the local optima that trip up the models. Under this hypothesis, no special additions are needed to explain the accuracy of human learners; the disparity is simply a result of insufficient computational methods.

Another possibility is that non-statistical methods can close the gaps. This is similar to the traditional definitions of bootstrapping; once a few words have been learned with high confidence, these words can be used to correct the meanings of other words. For instance, if *dog*
and chase are both well-established to mean dog and chase, but cat has been mistakenly mapped to child, hearing “Dog chases cat” while seeing a cat being chased might convince the learner to override the statistical learner and directly change the meaning of cat in the lexicon.

7 Future Work

The research undertaken in this paper is the first step in a much larger research program. The model, by setting up a joint-learning method under ambiguity and creating a measure for online expectations, lays the groundwork for an array of more extensive experiments.

Expectations in Language Learning One fertile field for further study is expectations in more complex learning environments. The artificial language learning paradigm can be used in many new conditions, and the results can be compared to the predictions made by the model. The original study by Amato and MacDonald looked at a simple learning task; learning a constant word-order language using unambiguous pictures. Both the adult language learners and our model can be tested on more complex acquisition tasks to get a better idea of how language acquisition proceeds in more complex situations and how well the model’s predictions fit with observed human learning. One direction to look at is the effect of ambiguous pictures during language acquisition. How do human learners deal with ambiguity? Do they switch their learning mechanism in any noticeable way? Does their behavior still follow the models’ predictions?

Another expansion is to look at how adults learn a language with variable word order. We have seen that the model can engage in probability matching behaviors because it keeps track of the relatively likelihoods of different syntactic structures. It is not entirely clear how humans treat syntactic variability in a language they are learning. By the time they attain fluency in it, they are able to use probabilistic information in their syntax, as seen in studies of gradient constraints on syntactic alternations, but are they able to use syntactic probabilistic information early in acquisition? The model predicts that they can, but this needs to be tested.

The monster-world paradigm is also essential for testing the hypothesis that non-linguistic data can be used to improve world knowledge, and thereby improve expectations in sentence processing. The model’s prediction that non-linguistic learning will occur can easily be checked by comparing human learners who encounter only sentence-picture pairs to others who see some captioned pictures and some uncaptioned ones. If this prediction holds true, there is also the empirical question of Section 6.3 — how should the model weight world knowledge gained in the absence of linguistic data relative to paired data?

Clear Hierarchical Structure Another extension of the model is to show that it can learn an unambiguously hierarchical syntactic structure. This can be done by adding a modifier to the language, such as an adjective or prepositional phrase, that modifies the head of the phrase that dominates the modifier. This becomes an especially interesting problem if the language being learned has VSO, VOS, SOV, or OSV word order, so that the two noun phrases are next to each other, and the modifier separates them. In that case, the learner must learn the headedness direction of the language and determine whether NP → N Adj or NP → Adj N is preferred in order to get the correct NP for the modifier. The model’s ability to learn which NP the modifier goes to can be tested both by looking at the relative likelihood of NP rewriting as N Adj or Adj N and by directly comparing the probability of the two different parse trees in a language where the two NPs abut. Extending the model to account for modifiers would also open the door to further expansions that could lead to modelling the acquisition of complex syntax, such as relative clauses, down the road.

Time Course of Acquisition At present, we are only looking at the grammar after the model has had a chance to settle down among grammars it feels comfortable with. This can take around 2000 iterations, but the Metropolis algorithm generates a grammar in each iteration. It could be enlightening to investigate the path by which the model reaches its final grammar to see if it behaves like a human or exhibits any of the difficulties that human language learners encounter. A possible comparison is to look at the rate at which humans and the model acquire rare words of a language.

Improved Semantics The model as presented uses a very simple semantic distribution over pictures to induce the observed online processing effects of second-order dependencies. More complicated processing effects can only be explained with improvements to the grammar of pictures, moving away from the simple multinomial used in this paper and into a more explanatory picture grammar. Another improvement to the semantics is the inclusion of a method of acquiring formal semantics. Future work with klinton Bicknell and Steve Piantadosi will work on these extensions of the model.

8 Conclusion

Our joint model of syntax and lexicon acquisition shows that statistical learning is powerful enough to overcome such issues as ambiguous inputs and a lack of negative evidence without relying on language-specific cognitive frameworks. The model learns a surprisingly accurate syntax and lexicon through cross-situational learning and statistical bootstrapping. The model’s framework is also useful after acquisition for studying psycholinguistic phenomena, such as predictions about upcoming words in a sentence. These predictions comply with data from adults learning an artificial language, suggesting that the model is on the right track.
Acknowledgments

Clinton Bicknell, Roger Levy, Steve Piantadosi, and Josh Tenenbaum were instrumental in working out the math behind the model and ironing out the kinks. Mike Amato and Maryellen MacDonald offered useful insight into the artificial language learning paradigm. Jamie Alexandre and Charles Elkan supplied helpful advice as well. This research was funded by an NIH training grant from the Center for Research in Language.

References


Doye, G. (2008). Determinants of variation in the needs doing construction. (First comprehensive paper, UCSD)


Gabriel Doyle  Learning Syntax and a Lexicon from Ambiguous Input  Comps Paper II

Cambridge: MIT Press.


Carlo with people. In Advances in neural information processing systems.


