# A Bayesian Model of the Acquisition of Compositional Semantics

**Steven T. Piantadosi, Noah D. Goodman, Benjamin A. Ellis, Joshua B. Tenenbaum**
{ **piantado, ndg, benjelli, jbt** } **@ mit.edu**
MIT Department of Brain and Cognitive Sciences
43 Vassar Street, Building 46, Room 3037, Cambridge, MA 02139

## Abstract

We present an unsupervised, cross-situational Bayesian learning model for the acquisition of compositional semantics. We show that the model acquires the correct grammar for a toy version of English using a psychologically-plausible amount of data, over a wide range of possible learning environments. By assuming that speakers typically produce sentences which are true in the world, the model learns the semantic representation of content and function words, using only positive evidence in the form of sentences and world contexts. We argue that the model can adequately solve both the *problem of referential uncertainty* and the *subset problem* in this domain, and show that the model makes mistakes analogous to those made by children. **Keywords:** Compositional semantics; language acquisition; Bayesian learning; Combinatory Categorial Grammar

## Introduction

Language is perhaps the only communicative system in nature which compositionally builds structured meanings from smaller pieces; this compositionality is the cognitive mechanism that allows for what Humboldt called language's "infinite use of finite means." A child acquiring language is therefore faced with learning both which words correspond to which objects and actions in the world (*lexical* semantics), and also how words can be pieced together to convey complex meanings (*compositional* semantics). This includes learning the formal semantic characteristics of content words that allow them to be used compositionally, and also inferring how function words transform abstract representations of meaning. For example, the word "every" in the sentence "Every person laughed" has an abstract meaning which quantifies the proposition $LAUGHED(x)$ over all people $x$. Because function words do not generally refer to entities in the world, but rather represent abstract syntactic and semantic operations, they provide an especially interesting learning problem.

Though computational models exist for many problems in language acquisition, the area of compositional semantics has been relatively unstudied. We present a novel, unsupervised, cross-situational learning model which is capable of learning compositional semantics in a relatively unconstrained hypothesis space. The model uses positive evidence in the form of sentences and contextual information from world environments to learn mappings from words to pairs of Combinatory Categorial Grammar (CCG) syntactic classes and lambda expressions. Together, the syntactic and semantic forms allow the model to compositionally build the meaning of sentences. The model contains an inductive bias to prefer grammars with shorter description lengths, and posits that the utterances it hears are typically true in its current context. We observe that the model faces two key problems in this domain: the *problem of referential uncertainty* and the *subset problem*. The problem of referential uncertainty concerns the fact that any sentence can potentially refer to an infinity of possible actions, objects, and logical relations. The subset problem is that learners may consider an under-restrictive grammar to be correct. If this happens, they will receive no direct evidence that contradicts their grammar, and may not acquire the correct one. We show that the model can solve these problems using a Bayesian cross-situational learning algorithm.

Language acquisition is a complex and multi-faceted problem, and for reasons of computational tractability, we focus only on compositional semantics. Thus, the model assumes that lexical semantics has already been acquired—it is given a mapping from words to objects or actions in the world. The model learns the semantic structures that govern how these words can combine in a toy version of English which includes nouns, quantifiers, transitive verbs, and intransitive verbs. The model also learns the semantic representations of words such as "every" which do not correspond to objects in the world, but provide consistent abstract operations on meanings. Focusing on only this dimension of the language acquisition problem allows significant simplification conceptually; however, this model could in principle be combined with similar models for learning lexical semantics from context (Frank, Goodman and Tenenbaum 2007). More generally, it could be combined with similar models for tasks such as concept learning (Goodman et al. 2008), word segmentation (Goldwater, Griffiths and Johnson 2006) and syntactic learning (Klein and Manning 2005; Perfors, Tenenbaum and Regier 2006) to provide a comprehensive statistical model of language acquisition.

In addition, the model accounts for *quantifier spreading*, a well-studied phenomenon in the quantifier acquisition. Inhelder and Piaget (1964) observed that children up to ages 4 and 5 often answer "no" to the question "Is every man wearing a hat?" in a situation in which every man is wearing a hat, but one hat is not being worn by anyone. A variety of explanations have been offered for this effect, children's misunderstanding of part-whole relationships (Inhelder and Piaget 1964), non-adult linguistic representations (Philip 1995), and pragmatic difficulties (Crain 2000). We show that our learning model makes analogous mistakes in learning, providing a statistical learning account of the effect.

### Previous research in compositional semantics learning

Models of compositional semantics acquisition have previously been proposed in natural language processing (Zettle-

moyer and Collins 2005, Wong and Moony 2007), robotics (Sugita and Tani 2003), and the description of visual scenes (Roy 2002). In general, these models would provide unsatisfactory psychological models due to their reliance on supervised learning, or implausible linguistic representations. Siskind (1996) proposed a cross-situational word learning algorithm as a psychological model, which employed an elementary compositional semantics. He showed how cross-situational inference can solve the problems of referential uncertainty, homonomy and noisy input.

We try to address several shortcomings of Siskind's model. First, Siskind's model treats sentences as unordered sets, and it is not clear how to extend his results to fit with contemporary syntactic theories. In contrast, this paper studies mappings from lexical items to lambda calculus expressions, which interface with many syntactic theories, including generative grammar and CCG (Steedman 2001). Siskind does not explicitly address the subset problem, and it is not clear that if his algorithm were extended to a learning domain for which the subset problem was a serious issue—such as the learning of quantifiers—it would perform adequately. In addition, Siskind's algorithm follows a set of seemingly ad hoc rules, which makes it difficult to extract specific learning principles which can be used to extend the work and combine it with other learning theories. In contrast, this paper does not attempt to provide an algorithmic theory, but rather shows how the compositional semantics can be acquired by any algorithm which solves the statistical problem we formalize below.

## Preliminaries

We choose to formalize syntax with CCG and semantics with the lambda calculus. These are used because CCG is a lexicalized syntactic formalism which easily interfaces with the lambda calculus. A lexicalized grammar is desireable because it simplifies the search problem: to find the best grammar, we must search over mappings of words to CCG syntactic types, and not separately learn rules of the grammar. While CCG generally makes use of type-raising rules, we use a simplified CCG grammar which does not require type-raising. Before introducing the model, we present the simplified versions of these formalisms used in the model.

A wide range of linguistic phenomena have been modeled using CCG, including coordination, WH-questions, prosody, and crossed-serial dependencies (Steedman 2001). In CCG, the syntactic category of each word specifies what syntactic types it requires to the left or to the right, and which syntactic type results from combining a word with its neighbor. For example, the word "loved" may be of type $(S\backslash NP)/NP$, where the outermost "/" denotes that "loved" requires an NP to the right and returns an element of type $S\backslash NP$. The type $S\backslash NP$ requires an NP to the left and returns a type $S$, for "sentence." In these expressions, "/" denotes a requirement to the right, and "\" denotes a requirement to the left. Thus, if "Mark" and "Jenn" are of type NP, the sentence "Mark loved Jenn" can be represented syntactically by $NP \;\; (S\backslash NP)/NP \;\; NP$. First, $(S\backslash NP)/NP$ and the final NP combine, giving $NP \;\; S\backslash NP$. Then, the first NP combines with $S\backslash NP$, giving a syntactic type S.

Many semantic formalisms—including CCG—represent word meanings as functions and constants. Semantic representations are well-integrated with the CCG syntax: any time two syntactic elements combine, their respective semantic functions are combined with functional composition. Thus, the syntax determines the order of composition[1]. The lambda calculus provides a convenient formalism for representing functions, and was originally studied in the context of computability and logic. In extremely simplified systems, for example, a verb such as "loved" may be mapped to the lambda expression

$$\lambda x \lambda y.LOVED(y,x). \tag{1}$$

Informally, (1) can be interpreted as a function which takes two arguments, $x$ and $y$, and returns the logical expression $LOVED(y,x)$. One can tell that $x$ and $y$ are arguments since they are each preceded by a $\lambda$.

The words "Mark" and "Sallie" may be mapped to the logical atoms *MARK* and *SALLIE* respectively. To arrive at the semantics of a sentence such as "Mark loved Sallie" we first compose the lambda expressions for "loved" and "Sallie," giving $\lambda y.LOVED(y,SALLIE)$. The result is then composed with the lambda expression for "Mark," giving the logical expression $LOVED(MARK,SALLIE)$, which is intended to be the logical representation of the meaning of "Mark loved Sallie."

In more complex linguistic constructions, the arguments to a lambda expression may be other lambda expressions. For example, in the sentence "Every man loved Sallie," a simplified analysis would map "man" to $\lambda y.MAN(y)$, and "every" to

$$\lambda x.\lambda y.\forall z.x(z) \rightarrow y(z). \tag{2}$$

Thus, "every" is mapped to a lambda expression that takes two arguments, $x$ and $y$, both of which are other lambda expressions. It returns the expression $\forall z.x(z) \rightarrow y(z)$, where $x(z)$ represents the lambda expression $x$ applied to $z$, and $y(z)$ represents the lambda expression $y$ applied to $z$. Thus, in the sentence "Every man loved Sallie," shown in Figure 1, we first compose "Every" and "man" to get $\lambda y.\forall z.MAN(s) \rightarrow y(z)$. We also compose "loved" and "Sallie" to get $\lambda y.LOVED(y,SALLIE)$. These expressions are then composed to give

$$\forall z.MAN(z) \rightarrow LOVED(z,SALLIE). \tag{3}$$

This expression represents the meaning of the sentence in that (3) is true if and only if the sentence "Every man laughed." is true. In this way, CCG and the lambda calculus can be used to define a first-order logical expression for any sentence, which then can be evaluated as true or false in a world context.

[1]Many semantic formalisms also use a type-system for this purpose.

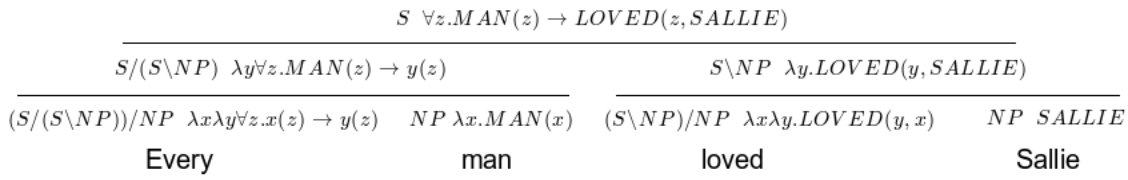


Figure 1: An example derivation showing syntactic CCG types and lambda calculus expressions combining to form a logical representation of the sentence "Every man loved Sallie."

### Problems in compositional semantics learning

As the amount of debate concerning the semantic structure of common words such as the definite article shows, the compositional semantic structure of language is neither obvious nor trivial. In addition to the sheer complexity of semantic representations children apparently acquire, learners also face impoverished input and a potentially infinite hypothesis space. Two problems in particular pose a substantial challenge to learners.

First, the problem of *referential uncertainty*—sometimes called the *gavagai problem*—concerns the fact that there is a large or potentially infinite set of meanings for each utterance a child hears (Quine 1960, Siskind 1996). Not only can each word potentially pick out an infinite set of objects in the world, each word may represent one a large set of possible logical meanings, or lambda expressions. For example, even if the child knows lexical semantics in that they know the word "loved" is some expression involving the logical or conceptual propositional $LOVED$, there are still many potential alternatives to the correct semantics (1) above. For example, the child might potentially consider expressions such as $\lambda x \lambda y.LOVED(x,y)$ or $\lambda x \lambda y.\forall z LOVED(x,z) \leftrightarrow LOVED(y,z)$. A well-formulated learning model must show how children could use observed sentences and contextual information to solve this problem.

Another key problem faced in compositional semantics learning is a variant of the *subset problem*. Like some models of grammar learning (Berwick 1985, Manzini and Wexler 1987) and phonological learning (Hayes 2004, Prince and Tesar 2004), it is possible that a learner would incorrectly infer an under-constrained hypothesis to be correct. For example, a possible logical representation of the sentence "Every man loved Sallie" is

$$\exists z.MAN(z) \rightarrow LOVED(z, SALLIE). \tag{4}$$

Note that the correct semantics (3) is logically stronger than (4). Therefore, a naive model which assumes the learner chooses a compositional semantics which maps sentences to meanings which are most often true would incorrectly choose a grammar which gives the sentence meaning (4). The learner must therefore have some mechanism to prevent them from learning grammars which are logically too weak.

## The model

We assume the task of the learner is to find the "best" way to map sentences that they hear into first-order logical forms. This mapping is formalized using a grammar $G$, which is a function from words to pairs of CCG syntactic types and lambda expressions. Once the learner has a CCG syntactic type and lambda expression for each word, we assume that they follow the syntactic and semantic composition rules to arrive at the representations of sentences in first-order logic. In addition, the model assumes that the sentences are uttered in world contexts and that the truth value of each logical expression can be evaluated in the world context.

Formally, let $A_1, A_2, \ldots$ be sets of sentences that are uttered in their respective world contexts, $C_1, C_2, \ldots$. Here, $C_i$ is taken to be a logical model of the world. We construct the $C_i$ by defining truth values for all atomic propositions such as $LOVED(BEN, JENN)$ and $CRIED(SALLIE)$, which then recursively defines truth values for combinations of these propositions with logical connectives and quantifiers. Like Siskind (1996), we assume that Quine (1960)'s problem is at least mitigated by some perceptual or conceptual system that narrows down the range of possible meanings to a salient set.

The task of the learner is then to find a mapping $G$ from words to pairs of lambda expressions and syntactic types that maximizes $P(G \mid C_1, C_2, \ldots, A_1, A_2, \ldots)$. We assume that each sentence is generated independently given a context and that contexts are independent of the grammar $G$. By Bayes rules, we therefore have,

$$\begin{aligned} P(G \mid C_1, C_2, \ldots, A_1, A_2, \ldots) &\propto P(A_1, \ldots \mid C_1, \ldots, G) P(G) \\ &= P(G) \prod_i P(A_i \mid C_i, G) \\ &= P(G) \prod_i \prod_{a \in A_i} P(a \mid C_i, G) \end{aligned} \tag{5}$$

We choose simple forms for both the likelihood $P(a \mid C_i, G)$ and the prior $P(G)$. Namely, we assume that with probability $\alpha$, the speaker chooses a true sentence uniformly at random, and with probability $(1-\alpha)$ the speaker chooses a false sentence uniformly at random. Thus we have

$$p(a \mid C_i, G) = \begin{cases} \frac{\alpha}{T_i} \text{ if } a \text{ is true} \\ \frac{1-\alpha}{F_i} \text{ if } a \text{ is false} \end{cases} \tag{6}$$

where $T_i$ is the number of sentences which are true in context $C_i$ according to $G$ and $F_i$ is the number that are false. For the

Table 1: Grammar for the target language

| Word | Syntactic type | Semantics |
|---|---|---|
| laughed, cried | $S\backslash NP$ | $\lambda x.LAUGHED(x)$ |
| loved, hated | $(S\backslash NP)/NP$ | $\lambda x\lambda y.LOVED(y,x)$ |
| Mark, Ben, Sallie, Jenn | $NP$ | $MARK$ |
| person, man, woman | $NP$ | $\lambda x.PERSON(x)$ |
| every | $(S/(S\backslash NP))/NP$ | $\lambda x.\lambda y\forall z.x(z)\rightarrow y(z)$ |
| some | $(S/(S\backslash NP))/NP$ | $\lambda x.\lambda y\exists z.x(z)\wedge y(z)$ |

simulations here, we fix $\alpha = 0.9$.

The prior $P(G)$ is defined to penalize complex grammars, which we take to be grammars which map words to lambda expressions containing many logical elements such as quantifiers, lambdas, atomic propositions, and logical connectives. In this way, the prior can be seen as a Minimum Description Length prior which penalizes grammars whose semantics require long expressions in the lambda calculus. Thus, if $\text{Cplx}(l)$ is the number of logical elements in the expressions $l$, and $G_L$ is a mapping from words to lambda expressions, we define

$$P(G) \propto \exp(-\sum_{w \in W} \text{Cplx}(G_L(w))),$$

where $W$ is the set of words in the language.

In addition, the model assumes that words of the same class are mapped to "similar" lambda expressions. That is, if the lambda expression for "loved" is $\lambda x\lambda y.LOVED(y,x)$, then the lambda expression for "hated" must be $\lambda x\lambda y.HATED(y,x)$, where the only change is that $LOVED$ has been replaced by $HATED$. This significantly reduces the size of the hypothesis space that needs to be considered and is generally true in semantic theories (Steedman 2001). Moreover, these word classes are plausibly learned by other means, such as distributional analysis (Cartwright and Brent 1996; Redington, Chater and Finch 1998; Mintz, Newport and Bever 1998).

Thus, in order to maximize (5), the learner must find a "simple" grammar which assigns high likelihood to the sentences heard. To solve this computational problem, we employ a backtracking search that finds the exact maximum for the hypothesis space defined above[2]. We prune the search by not considering grammars that give malformed syntactic or semantic expressions, thereby assuming that the learner only considers grammars which can perfectly parse all of the sentences heard. This assumption is not realistic, but simplifies the search problem considerably.

## The target grammar and model implementation

We work with a toy version of English which, though simple, still faces the subset problem and problem of referential uncertainty discussed above. The target grammar is shown in Table 1 and possible sentences are of the following forms: intransitives ("Mark smiled"), transitives ("Mark loved Sallie"), subject-quantified intransitives ("Every boy smiled"), and subject-quantified transitives ("Every boy loved Sallie").

[2]We choose at random when grammars tie.

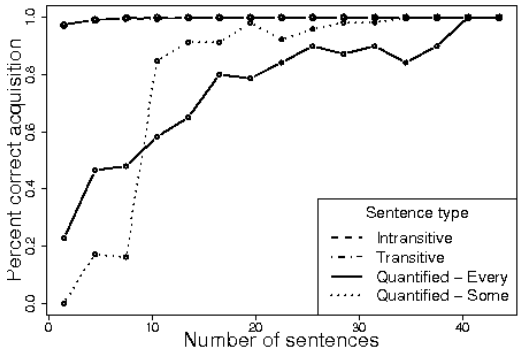


Figure 2: Probability of learning a grammar which gives the correct meaning for each sentence type, plotted against the number of sentences of that type seen in the input ($K = 0.15$).

As above, we assume that words in each row of Table 1 are mapped to "similar" lambda expressions. In general the number of possible grammars grows exponentially in the number of words; therefore for computational tractability we make several simplifying assumptions in the model implementation:

*The lexical semantics is already known by the model.* We take this to mean that the model knows the lambda expression for a word such as "loved" may or may not involve the logical element $LOVED$, but definitely *will not* involve the logical elements $HATED$ or $CRIED$. Again, we believe this is a reasonable assumption because other methods such as Frank, Goodman, and Tenenbaum (2007) have shown that lexical semantics can be learned by a similar model. Though this mitigates the problem of referential uncertainty for the lexicon, there is still considerable uncertainty about what logical relations in the world sentences refer to.

*Sentences without quantifiers are learned before sentences with quantifiers.* This considerably shrinks the hypothesis space: we first search over lambda expressions and CCG syntactic types for proper nouns ("Mark," "Ben," "Sallie," "Jenn"), intransitives ("laughed" and "cried") and transitives ("loved" and "hated"). We then fix these and search over lambda expressions for lambda expressions and CCG syntactic types for common nouns ("Person," "man," "woman"), "every," and "some." This assumption is empirically justified in that children do not typically learn quantifiers until long after they have learned simple verbs (Inhelder and Piaget 1964).

*The space of lambda expressions considered by the model is finite but large.* This spaces includes all expressions with less than or equal to two function applications, one logical connective ($\rightarrow, \leftrightarrow, \vee, \wedge, \neg$), one quantifier, and two bound lambda variables. In addition, we removed lambda expressions which contained unused variables bound by quantifiers or lambdas, resulting in a total of 2657 possible lambda expressions for each word.

However, the computational task faced by this model is still formidable. For both sentences with quantifiers and sentences without quantifiers, the model considers over *a trillion* poten-

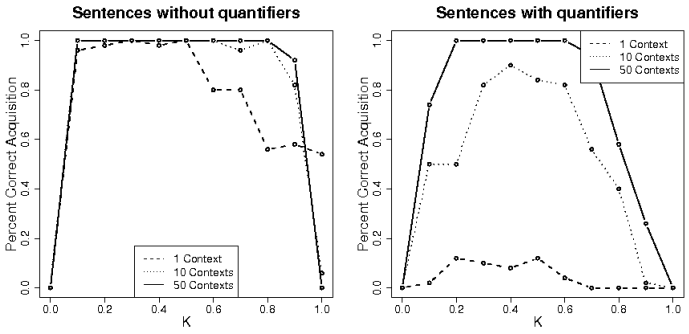


Figure 3: Probability of learning the correct mapping from sentences to logical forms for non-quantified sentences and quantified sentences for various $K$ given 10 sentences per context. Each point represents the an average of 50 runs of the model.

tial grammars, many of which are eliminated by backtracking. As far as we know, such as large hypothesis space has not been considered previously in this domain, and we show that even an unsupervised model is capable of learning the correct grammar. We expect that the model could be scaled up to larger grammars by using appropriate algorithms.

## Results and discussion

The model we present learns compositional semantics in the form of lambda expressions and CCG syntactic types, which are used to map sentences to first-order logical expressions. The model assumes that the sentences are typically true in the world context, and attempts to find a simple grammar which assigns the observed data high likelihood. However, because we know of no data set which gives both the contexts and sentences that occur in them, we created such a data set ourselves. One free parameter, $K$, governs what proportion of atomic propositions such as $SMILED(BEN)$ and $LOVED(JENN, MARK)$ are true on average in each context. Intuitively, acquisition will be difficult or impossible when $K$ is close to 0 or 1, since in these instances the world context will not be informative about the meanings of sentences.

Figure 2 shows how often the model acquires the correct semantic mapping for sentences with transitive verbs, intransitive verbs, "every" and "some" as a function of how many sentences it is given containing each of these words. Note that transitive and intransitive verbs are acquired very quickly—often after only a few sentences containing them. This is because their semantics are short and therefore favored by the prior. The correct semantics for sentences containing "every" and "some" is also learned quickly, typically after about 20 sentences. Interestingly, though, the semantics of "every" is learned less quickly than "some."

Though it is not clear what an ecologically valid value for $K$ is, we show that the model can acquire the correct semantics over a wide range of values for $K$. Figure 3 shows the probability that the model acquires a mapping which gives the correct meaning, as a function of $K$ and the number of contexts the model has seen. Note that the model quickly learns to map sentences to the correct logical forms for a wide range of $K$ values—often after 1 context for transitive and intransitive verbs, or 50 contexts for quantifiers. It has also been suggested that rarity of true propositions (corresponding to small $K$) is an important property in reasoning and learning (Oaksford and Chater 1994). Indeed, Figure 3 shows an asymmetry, learning more quickly when $K$ is small than when $K$ is large.

We also analyzed the mistakes that the model makes late in acquisition. Table 2 shows probability with which the model assigned the sentence "Every woman laughed" various semantic values, for $K = 0.1, 0.8, 0.9$ after 50 contexts[3]. For other values of $K$ the model always learned the correct semantics. We note that for $K = 0.1$, the only mistake the models makes is to interpret the sentence "Every woman laughed" as being true if and only if all women laughed, and the only people who laugh are women. This is analogous to the mistake children make in quantifier spreading. From a learning perspective, this mistake is reasonable because the biconditional and implication operators are logically very similar, and therefore require specific and potentially rare kinds of data to distinguish. For $K = 0.8$ and 0.9, the model often chooses $\forall z.LAUGHED(z)$, ignoring the restriction to women. Similar mistakes have been reported in children (Kang 1999), but it is likely that typical $K$ values are much lower (Oaksford and Chater 1994)

| | $K$ | | |
|---|---|---|---|
| **"Every woman laughed."** | **0.1** | **0.8** | **0.9** |
| $\forall z.WOMAN(z) \rightarrow LAUGHED(z)$ | 0.74 | 0.98 | 0.62 |
| $\forall z.WOMAN(z) \leftrightarrow LAUGHED(z)$ | 0.26 | 0 | 0 |
| $\forall z.LAUGHED(z)$ | 0 | 0.02 | 0.36 |
| $\forall z.LAUGHED(z) \vee WOMAN(z)$ | 0 | 0 | 0.02 |

Table 2: Example mistakes made in learning

As discussed above, unsupervised learners of compositional semantics face two primary problems: the problem of referential uncertainty, and the subset problem. In this model, the cross-situational aspect of the learning allows the problem of referential uncertainty to be solved. The optimal grammar is the one which performs best across a wide range of contexts, so even though individual sentences could refer to a range of possible logical relations, the model chooses a grammar which consistently produces highly-probable meanings across all contexts it sees.

The subset problem is solved in this model by the *size principle* (Tenenbaum 1999). Because the model assumes that

[3]These probabilities are the average of 50 model runs.

| | Rank | Log Posterior | LL | Log Prior | "man" | "every" | "some" |
|---|---|---|---|---|---|---|---|
| **(a)** | 1 | -1716 | -1698 | -18 | $\lambda x.MAN(x)$ | $\lambda x \lambda y \forall z.x(z) \rightarrow y(z)$ | $\lambda x \lambda y \exists z.y(z) \wedge x(z)$ |
| **(b)** | 2 | -1723 | -1705 | -18 | $\lambda x.MAN(x)$ | $\lambda x \lambda y \forall z.y(z) \leftrightarrow x(z)$ | $\lambda x \lambda y \exists z.y(z) \wedge x(z)$ |
| **(c)** | 731 | -2035 | -2017 | -18 | $\lambda x.MAN(x)$ | $\lambda x \lambda y \exists z.x(z) \rightarrow y(z)$ | $\lambda x \lambda y \exists z.y(z) \wedge x(z)$ |

Table 3: Key semantic values for $K = 0.15$ with 50 contexts. These grammars all have the correct CCG syntactic types.

true sentences are chosen from a probability distribution, it will be penalized if the grammar maps sentences to logical forms which are true too often[4]. That is, the likelihood of all sentences will be lowered if the model posits incorrectly that some sentences are true. Table 3 illustrates this with the results of a typical run. The table shows the rank of the grammar's overall posterior relative to all other grammars, as well as the posterior, prior, and likelihood. Note that (a), the highest scoring grammar, is the correct one on this run of the model. The second highest scoring grammar is (b), which uses the biconditional rather than implication in the lambda expression for "every." The fact that these are close and highly-ranked is consistent with Table 2, which shows that the model often chooses the wrong one, making the same mistake as children. In addition, (c) is ranked relatively low. This is interesting because (c) is the logically weaker version of the correct grammar that, and thus would be preferred by a learner which chooses grammars which are most often true.

## Conclusion

We have shown that an unsupervised, Bayesian model can correctly learn compositional semantics and syntax in a toy version of English, using only positive evidence and contextual information. Moreover, this model considers a relatively unconstrained hypothesis space, and makes mistakes in acquiring quantifiers analogous to those children make. We have argued that the toy language presents some of the core problems that children face in acquiring a full compositional semantics. The model succeeds due to very general principles in Bayesian learning and uses linguistically-realistic representations of semantics. This suggests that the model could be scaled up to include more realistic models for sentence production, or a more elaborate syntax and semantics.

## References


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

Cartwright, T., & Brent, M. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, *63(2)*, 121–170.

Crain, S. (2000). Sense and sense ability in child language. In *Proceedings of the 24th annual boston university conference on language development.* Somerville: Cascadilla Press.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A bayesian framework for cross-situational word-learning. In *Advances in Neural Information Processing Systems 20.*

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of Coling/ACL.*

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*.

Hayes, B. (2004). Phonological acquisition in optimality theory: The early stages. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Fixing priorities: Constraints in phonological acquisition.* Cambridge, UK: Cambridge University Pres.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. London: Routledge.

Klein, D., & Manning, C. (2005). Natural language grammar induction with a generative constituent-context model. *Pattern Recognition*, *38(9)*, 1407-1419.

Manzini, R., & Wexler, K. (1987). Parameters, binding theory, and learnability. *Linguistic Inquiry*, *18*, 413–444.

Mintz, T., Newport, E., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, *26(4)*, 393–424.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2006). Poverty of the stimulus? A rational approach. In *Proceedings of the cognitive science society.*

Philip, W. (1995). *Event quantification in the acquisition of universal quantification*. Ph.D. thesis, University of Massachusetts, Amherst.

Prince, A., & Tesar, B. (2004). Learning phonotactic distributions. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Fixing priorities: Constraints in phonological acquisition* (pp. 41–76). Cambridge, UK: Cambridge University Press.

Quine, W. V. (1960). *Word and object*. Cambridge, MA: MIT.

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

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61(1-2)*, 31–91.

Steedman, M. (2001). *The syntactic process*. Cambridge, MA: MIT Press.

Tenenbaum, J. (1999). *A bayesian framework for concept learning*. Ph.D. thesis, MIT.


[4]The distribution need not be uniform for this to hold.