Sampling assumptions affect use of indirect negative evidence in language learning

Anne S. Hsu
Department of Cognitive, Perceptual and Brain Sciences
University College London

Thomas L. Griffiths
Department of Psychology
University of California, Berkeley

Address for correspondence:
Anne Hsu
26 Bedford Way,
London, WC1H 0DS
United Kingdom
Email: ahsu@gatsby.ucl.ac.uk
Abstract

A classic debate in cognitive science revolves around understanding how children learn complex linguistic rules, such as those governing restrictions on verb alternations, without negative evidence. There are two formal perspectives on language acquisition, which differ in their assumptions about how the problem of learning a language is formulated. In discriminative learning, the goal is to learn to classify sentences as grammatical or ungrammatical. Here, the probability distribution from which the sentences are sampled is irrelevant. In generative learning, the goal is to estimate a probability distribution over sentences. These assumptions have strong implications for language learning: Not hearing a particular sentence produced provides indirect negative evidence against that sentence being in the language only under the assumptions of generative learning. We demonstrate that human learners can produce behavior consistent with the predictions of both approaches, depending on how the learning problem is presented. These results suggest that people use information about the way in which linguistic input is sampled to guide their learning.
Sampling assumptions affect use of indirect negative evidence in language learning

How children acquire language has been the focus of a major debate in cognitive science for over 50 years (e.g., Chomsky, 1959). A key question in this debate is how children learn the correct generalizations of grammatical rules. Successful linguistic communication requires the ability to generalize beyond the utterances one has heard, in order to form new utterances. However, the rules of a language often have exceptions, creating many opportunities to form generalizations that are ungrammatical. One such example is exceptions on the dative verb alternation in English. Many verbs, such as *give*, undergo the dative alternation, which means that they appear in both the direct construction, *I gave her the hat*, and the prepositional construction, *I gave the hat to her*. However, the verb *donate* is an exception to this. *Donate* cannot undergo the dative alternation, and is only grammatical in the prepositional construction: *I donated a hat to the shop* is grammatical whereas *I donated the shop a hat* is not. The ability to learn such exceptions to generalizations presents a puzzle in light of the observation that children learn language from receiving mostly positive input, i.e. being exposed to grammatical utterances, with comparatively little feedback on what is ungrammatical (Bowerman, 1988; Brown & Hanlon, 1970; Marcus, 1993). So, how do children learn which generalizations are allowed and which ones are not?

Mathematical analyses of language learning have played a key role in this debate, producing results that have been used to argue both for and against the idea that innate domain-specific constraints guide language acquisition. These analyses have taken two distinct approaches to framing the language acquisition problem. One approach frames language acquisition as the problem of learning to identify grammatically acceptable and unacceptable sentences (Baker, 1979; Chomsky, 1965; Crain & Lillo-Martin, 1999; Gold, 1967; Pinker, 1989). This effectively means learning a mapping between sentences and
labels of grammaticality. The other approach frames language acquisition as the problem of estimating the probability distribution over grammatical sentences (Dowman, 2000; Langley & Stromsten, 2000; Onnis, Roberts, & Chater, 2002; Perfors, Regier, & Tenenbaum, 2006; Perfors, Tenenbaum, & Wonnacott, 2010; Stolcke, 1994). The key difference between these two approaches is the assumptions made about how the input is sampled. The first perspective, in which one learns a mapping between sentences and grammaticality, does not make any assumptions about the distribution from which the sentences are drawn. The second perspective, in which one learns a probability distribution over grammatical sentences, assumes that the linguistic input is sampled from the distribution associated with the language.

As a consequence of their different assumptions about how linguistic input is generated, these two approaches differ in their use of indirect negative evidence.\(^1\) While explicit feedback that a sentence is ungrammatical is relatively rare, indirect negative evidence can potentially be obtained from the observation that a sentence does not appear in the input (Braine, 1971). Learning a mapping from sentences to grammaticality does not require assumptions about the distribution from which the observed sentences are drawn, so an absent sentence does not give any evidence towards its ungrammaticality. On the other hand, if we assume that sentences are generated from the probability distribution over sentences associated with the language being learned, an absent sentence is probably not allowed in the language. Using indirect negative evidence in this way has been argued to provide an explanation of how children might learn language without requiring strong innate constraints (Rohde & Plaut, 1999; Chater & Manning, 2006).

These two perspectives on language acquisition map well onto the distinction between generative and discriminative models used in machine learning and

---

\(^1\) Also sometimes referred to as “implicit negative evidence” (e.g., Rohde & Plaut 1999).
computational linguistics. In the context of language, generative models aim to learn the probability distribution from which the observed linguistic input was generated. These models have recently been successfully used to model the acquisition of many aspects of language (for a review, see Chater & Manning, 2006). In contrast, discriminative models aim to solve a classification problem, learning a mapping from linguistic input to class labels (e.g., Roark, Saraclar, Collins, & Johnson, 2004; Collins & Koo, 2005; Petrov & Klein, 2008). Typically, no assumptions are made about the distribution from which the input is generated. Under this approach, learning a language reduces to learning to classify sentences as grammatical or ungrammatical.

The key difference between generative and discriminative approaches to language learning comes down to the assumptions that are made about how linguistic input is sampled. The implications that these assumptions have for the use of indirect negative evidence provide an opportunity to explore whether human learners are sensitive to sampling assumptions in learning the structure of a language. Previous work on word learning shows that both adults and children are sensitive to the way in which examples of a word are generated: generalization is reduced when examples are assumed to be generated by sampling from the set of referents of the word, rather than by labeling randomly selected objects (Xu & Tenenbaum, 2007a, 2007b). These two forms of sampling are analogous to the difference between the assumptions of generative and discriminative approaches to grammar learning. The reduction in generalization seen when examples are sampled from the set of referents is similar to using indirect negative evidence. However, it is an open question whether people show the same sensitivity to sampling in other aspects of language learning.

In this paper, we examine whether people are sensitive to sampling assumptions when learning an artificial language containing alternating and non-alternating verbs. The linguistic input contains examples of grammatical and ungrammatical constructions in our artificial
language. However, there is one construction that never appears in our training input. This missing construction is analogous to a restricted verb alternation that never appears in the language (such as donate in the direct dative construction). Thus we will be testing whether the absence of this construction for this verb is used as indirect negative evidence, i.e. whether people judge it to be ungrammatical because it is absent from the input. We first use computational modeling to show how generative and discriminative models differ in their use of indirect negative evidence with this artificial language. We then conduct two experiments that explore whether people are sensitive to the different sampling assumptions made by these different learning approaches, using their judgments of the grammaticality of the missing construction to identify whether they are using indirect negative evidence.

Models of learning: Generative and discriminative

The distinction between generative and discriminative learning has its origin in machine learning and statistics (Efron, 1975; Ng & Jordan, 2001; O'Neill, 1980; Press & Wilson, 1978; Ruiz-Velasco, 1991; Xue & Titterington, 2008). Traditionally, these modeling approaches have been applied to the problem of category learning (known as classification in the machine learning literature). The problem of categorization can be formulated in terms of estimating the probability an input $x$ belongs to a category $c$, $p(c|x)$. The difference between generative and discriminative learning approaches comes down to how they estimate this probability. Discriminative models directly estimate $p(c|x)$, learning the mapping between inputs and category labels. Generative models build a probabilistic model of the input by learning the probability that an input $x$ is generated given that the category is $c$, $p(x|c)$, and then solving the categorization problem by applying Bayes’ rule, with $p(c|x) = p(x|c) p(c) / p(x)$. Generative models assume that each input is sampled from the distribution associated with the corresponding category, $p(x|c)$. In contrast, discriminative models do not make any
assumptions about the distribution from which the input is sampled, since \( p(c|x) \) can be estimated without needing to make assumptions about the distribution of \( x \).

The performance of generative and discriminative approaches to category learning have been compared extensively (Efron, 1975; Ng & Jordan, 2001; O’Neill, 1980; Press & Wilson, 1978; Ruiz-Velasco, 1991; Xue & Titterington, 2008). This has typically been done by investigating a pair of models that are mathematically equivalent in every respect except the way in which their parameters are estimated. The classic \textit{discriminative-generative pair} being compared is usually (discriminative) logistic regression and (generative) naïve Bayes, which both assume that inputs \( x \) have a set of features that independently influence the probability of belonging to a category. Mathematical analyses and computer simulations show that generative and discriminative models exhibit significantly different category learning behavior. For example, if the training data consists of two normally distributed samples, generative models learn categories more quickly (Efron, 1975; Ng & Jordan, 2001; O’Neill, 1980). However, when the training data comes from other distributions, discriminative models are asymptotically more accurate (Press & Wilson, 1978; Xue & Titterington, 2008), though in some cases generative models may perform better initially and arrive at their (higher) asymptotic error more quickly (Ng & Jordan, 2001).

The distinction between generative and discriminative models can also be applied to language acquisition, where the input \( x \) is a sentence and the categories \( c \) are grammatical and ungrammatical sentences. The generative model estimates the probability distribution over sentences in the language, \( p(x|c = \text{“grammatical”}) \), and assumes that the input is sampled from this distribution. The discriminative model simply learns a function mapping each sentence to a grammaticality value, \( p(c = \text{“grammatical”}|x) \). Since the discriminative model makes no assumptions about the
distribution of the sentences and how this distribution relates to grammaticality, the absence of a sentence from the input does not influence whether it is considered grammatical. Consequently, we should expect that these two approaches should differ in how they handle indirect negative evidence. In the remainder of this section, we conduct a computer simulation to verify that this is the case.

An artificial language

To examine the predictions that these two models make about the use of indirect negative evidence we used an artificial language containing verb alternations. Our artificial language is based on one used in past work, which showed that adults are capable of using indirect negative evidence to learn restrictions on verb alternations in certain probabilistic language contexts (Wonnacott, Newport, & Tanenhaus, 2008; Perfors et al., 2010). Our language consists of three-word sentences, each containing a subject (N1), object (N2) and verb (V), with the order depending on the particular sentence structure. This language has four transitive verbs. Three of the four verbs are grammatical in only one sentence structure (non-alternating), while one of the four verbs is grammatical in two possible sentence structures (alternating). The status of the fourth verb is ambiguous, with insufficient evidence being presented to determine whether or not it alternates. We refer to this as the exception verb.

Vocabulary. The vocabulary for our language is a subset of that used in Wonnacott et al. (2008). There are three two-syllable nouns, each beginning with a different consonant, referring to three cartoon animals: blergen (lion), nagid (elephant), tombat (giraffe). Noun referents are fixed across participants. There are four one-syllable verbs: gund, flern, semz, and norg, corresponding to the four transitive actions: eclipse, push-to-side, explode and jump-on. While the meanings of the nouns and verbs are irrelevant to the models, we developed this language with the intent of also
examining human learning, as described below.

**Syntax and grammar:** In our language of three-word sentences, a verb could appear in three different positions (as the first, second or third word). We constrained the possible sentences such that the subject, N1, always appeared before the object, N2. This leaves us with three possible sentence structures, S1, S2, and S3, each of which corresponded to one of the following word orders: N1-N2-V, N1-V-N2 and V-N1-N2. In our experiment, the mapping from sentence structure to word order was randomized among participants. For example, S1 might correspond to N1-N2-V for one participant or it might correspond to V-N1-N2 for another participant. There was always one sentence structure, which we denote S3, that was never grammatical for any of the verbs. For S1 and S2, grammaticality varied depending on the verb. The language contained three verbs. One of the three non-alternating verbs was only grammatical in S1. The other two non-alternating verbs were only grammatical in S2. For example, let's consider the situation where S1 is N1-V-N2, S2 is N1-N2-V and S3 is V-N1-N2. If flern was an alternating verb, both nagid flern tombat and nagid tombat flern. would be allowed. If semz was non-alternating, and only allowed in S2, nagid tombat semz would be grammatical and nagid tombat semz would be ungrammatical. In this example, flern nagid tombat and semz nagid tombat are both ungrammatical.

**Learning the exception verb.** We can investigate how generative and discriminative learners handle indirect negative evidence by examining their treatment of the exception verb, which is never shown in S2. The crucial question is whether learners (model and human) categorize the missing construction as ungrammatical or grammatical in S2, meaning they either were or were not using indirect negative evidence, respectively. The frequencies with which positive and negative evidence regarding the four verbs and three sentence structures were presented are shown in Table
1. Grammatical and ungrammatical sentences are indicated with “+” and “−”, respectively, while “?” indicates that grammaticality is underdetermined by the data. The number in parentheses is the frequency with which each sentence (and accompanying information about its grammaticality) was presented. Verb V4 was never shown in sentence structure S2, making it possible for grammaticality predictions for sentences containing this verb to be used to explore the interpretation of indirect negative evidence.

Table 1: Artificial language used in initial simulations and Experiment 1

<table>
<thead>
<tr>
<th>Verb</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>+ (9)</td>
<td>+ (9)</td>
<td>- (6)</td>
</tr>
<tr>
<td>V2</td>
<td>- (3)</td>
<td>+ (18)</td>
<td>- (3)</td>
</tr>
<tr>
<td>V3</td>
<td>+ (18)</td>
<td>- (3)</td>
<td>- (3)</td>
</tr>
<tr>
<td>V4*</td>
<td>+ (18)</td>
<td>? (0)</td>
<td>- (6)</td>
</tr>
</tbody>
</table>

Note: +/−/? indicates grammaticality of each combination of verb and sentence structure, numbers in parentheses indicate frequency of presentation.

**Generative and discriminative models**

To determine the consequences of taking different approaches to learning the grammaticality of these sentences, we compare the predictions of a (generative) hierarchical Bayesian model (HBM) and a (discriminative) logistic regression model. We chose an HBM as our model of generative learning because it has previously been proposed as a model of a similar task, being a simplification of a more general grammar-learning approach (Perfors et al., 2010). Other generative models, including models that estimate full probabilistic grammars would behave similarly, and would also set the probability of an absent sentence near zero. For our discriminative model, we chose logistic regression
because it is simple and widely familiar to the statistics non-expert. We emphasize that we do not claim that these are the definitive models representing these two different approaches to formalizing the problem of learning a language. These are just simple models that are commonly used, which suffice to prove our point. Our modeling results are not tied to the specific models we chose. Other generative models such as probabilistic grammars (e.g., de Marcken, 1996) and discriminative models such as kernel machines (e.g., Jäkel, Schölkopf, & Wichmann, 2009) will behave similarly in their treatment of indirect negative evidence for the reasons mentioned above.

**Generative hierarchical Bayesian model.** For our generative model we use a hierarchical Bayesian model (HBM) specified in Perfors et al. (2010). The HBM allows for the learning of both verb-specific regularities as well as general abstract knowledge: A learner can learn the constructions associated with each verb, and form higher-level generalizations about how constructions tend to be distributed over the different verbs. For example, they could learn that all verbs are grammatical in only one possible construction, or that verbs tend to alternate, i.e. be grammatical in two possible constructions. The verb-general knowledge can also provide knowledge about the proportion in which constructions would be expected to appear. For example, one could learn that all verbs appeared in two possible constructions and tend to appear in the first construction 80% of the time. In this HBM, the general-level knowledge is represented probabilistically. For example, with two possible constructions, the general-level knowledge would be a probability distribution over how likely verbs are to appear in the first construction (as opposed to the second).
Figure 1: A hierarchical Bayesian model. This figure is adapted from Kemp et al. (Kemp et al., 2007)

A graphical depiction of an HBM is shown in Figure 1 (adapted from Perfors et al., 2010). Here, item-specific knowledge (also known as Level 1 knowledge) about the proportion of the different constructions that occur with each verb is represented by $\theta$. This value, $\theta$, is estimated based on verb-general knowledge (also known as Level 2 knowledge), which is represented in the HBM by two parameters, $\alpha$ and $\beta$. $\alpha$ represents the extent to which different verbs tend to occur with only a single construction and $\beta$ captures the occurrence probabilities of each construction, averaged over all verbs. The verb-general knowledge, in turn, depends on a third level of verb-general knowledge. This Level 3 knowledge is represented by the hyper-parameters $\lambda$ and $\mu$, which are prior distributions on the values of $\alpha$ and $\beta$, respectively. $\lambda$ determines the distribution over distributions of $\alpha$, i.e. distribution over the distributions determining uniformity. $\mu$ determines the distribution over distributions of $\beta$, i.e. distribution over the distributions determining the construction frequency over all verbs.

When applying an HBM model to data, inferences can be made at all of these different levels, based on the statistics of the input. In this case, the input consists of a set of sentences, each of which features a particular verb occurring in a particular
construction. Thus the input will contain a set of verbs and the distribution of
constructions in which each verb occurred. From this input, we can infer the values $\alpha$, $\beta$, $\lambda$, and $\mu$. In practice, for our purposes, we will fix $\lambda$ and $\mu$, which is equivalent to assuming that Level 3 knowledge is known. Hence we will only estimate the values of $\alpha$ and $\beta$ (see Appendix for further details).

**Discriminative logistic regression model.** For our discriminative model we use logistic regression. A logistic regression model can be used to learn a function that classifies observations into two classes. In the context of language learning, the observations are sentences and the classification problem is deciding whether each sentence is grammatical, as introduced previously. Logistic regression, like other discriminative models, takes in both positive and negative examples. Logistic regression learns a function that takes in a vector representing a data item and returns a probability of belonging to a particular class. Sentences will be represented in a vector format, which will represent the verb used, the particular construction in which the verb appears, and the combination of verb and construction specific to that sentence (see Appendix for details). Each vector will be associated with a class label, being either grammatical or ungrammatical. Based on the training set of input sentences, a function is learned that maps sentences in vector representation to a probability distribution over labels of grammatical or ungrammatical.

**Modeling results**

We compared how the two types of models differed in their predictions regarding the grammaticality of different sentences based on the input summarized in Table 1, and in particular how they would classify the exception verb, V4 in S2. The results of training the HBM on the grammatical sentences in the input show that V1 is expected to occur in both S1 and S2 50% of the time, as is the case in our language input. All other
verbs are expected to occur 100% of the time in the one grammatical sentence structure observed in the input. Predictions for grammaticality were extracted from the HBM model as follows: The sentence is grammatical if the probability of sentence structure $j$ for verb $i$, $\theta_j^i$, is greater than a pre-set parameter $\epsilon$ and ungrammatical when $\theta_j^i$ is less than $\epsilon$, corresponding to the assumption that grammaticality judgments are made by comparing the distribution over sentences estimated by the model with a uniform distribution over $1/\epsilon$ sentences. In our simulations, $\epsilon$ was set to be smaller than the value that $\theta_j^i$ would take if even one observed sentence used a particular structure. Grammaticality results for this model are shown in Figure 2.

![Grammaticality results for the HBM model](image)

**Figure 2.** Model predictions for grammaticality judgments from generative and discriminative models. The exception verb V4 is never shown in S2.

For the discriminative model, logistic regression was performed on the linguistic input. Occurrence frequencies of grammatical sentence structures in the input were the same as for the generative model. In contrast to the generative model, the discriminative model also takes negative input (examples of ungrammatical sentences), and these were presented with the frequencies shown in Table 1. Verbs V1-V3 are ungrammatical in two sentence structures, and each of these ungrammatical structures occurred three times in
the input. The exception verb V4 was never shown in S2. Thus in order to match the presentation frequency to that of other verbs, V4 appeared six times in the ungrammatical structure S3. Each verb-sentence structure observation was accompanied by the appropriate label of grammatical or ungrammatical. Predictions for grammaticality from the logistic regression model corresponded to the probability that sentence \( i \) is grammatical as computed by the model, \( p(c_i = \text{“grammatical”} \mid x_i, \eta, b) \) (see Appendix). Results are shown in Figure 2.

Inspection of Figure 2 shows that the predictions of the models regarding grammaticality are similar for all verbs and sentence structures except the critical case of V4 in S2. Here, the generative model indicates that this construction is ungrammatical, and the discriminative model leans towards grammaticality. The difference between the generative and discriminative modeling results can be understood as follows: A generative model learns a distribution over grammatically allowed sentences. In fact, it does not use negative examples (i.e. examples of ungrammatical sentences) in learning. As a consequence, when the exception verb never appears in S2, a generative learner receives indirect negative evidence and assumes it is not part of the language. On the other hand, the discriminative learner assumes that the grammatical examples are not generated from any distribution. Thus, when the exception verb never appears in S2, a discriminative learner receives no direct information about its grammaticality, and instead based on the other sentences it has learned. Since S2 and V4 both tend to be grammatical when viewed across all sentences in the language, and the model makes predictions based on features corresponding to the verb and sentence structure being used, the model predicts that S2 is likely to be grammatical for V4.

**Experiment 1: Human learning**

The simulations presented in the previous section illustrate how generative and discriminative approaches to language learning differ in their treatment of indirect negative
Sampling assumptions in language learning

This raises the question of whether a similar difference can be produced in human learners by changing the sampling assumptions appropriate for a language learning task. To explore this possibility, we conducted an experiment in which participants learned the artificial language used to generate the model predictions from the previous section, and manipulated the information participants received about how the linguistic input was generated.

Participants learned the artificial language by watching computer animated scenes accompanied by spoken and written sentences describing each scene. Participants were also provided with information about whether the sentence was grammatical or ungrammatical. Participants were assigned to one of two conditions, which prompted either generative or discriminative learning. Participants in both conditions were exposed to exactly the same sentences and grammaticality information. The two conditions differed only in how grammaticality information was presented.

**Method**

**Participants.** A total of 22 participants (11 in each condition) were recruited from the University of California, Berkeley community. Participants were compensated at a rate of $12/hour.

**Stimuli.** Participants were presented with both grammatical and ungrammatical examples of sentences from the artificial language described in Table 1. Noun referents were fixed across participants while the mapping of verbs to actions was randomly selected for each participant. As summarized in Table 1, participants viewed each of the 4 verbs 24 times, 18 grammatical sentences and 6 ungrammatical sentences. The alternating verb was shown 9 times each in S1 and S2 and 6 times in S3. The non-alternating verbs were shown 18 times each in their respectively grammatical sentence structures and 3 times each in the 2 ungrammatical structures. Presentation of sentences was ordered as follows: Two chains of
sentences were constructed, one grammatical and one ungrammatical. The grammatical chain consisted of 72 sentences (18 for each verb) and the ungrammatical chain consisted of 24 sentences (6 for each verb). For each sentence chain, verbs were presented cyclically and randomized within cycles. For the grammatical chain, V1 occurrences of S1 and S2 were cycled through in semi-random order (verbs V2-V4 appeared grammatically in only one sentence construction). Similarly, for the ungrammatical chain, V2 and V3 cycled semi-randomly through occurrences of S1 and S3 and S2 and S3 respectively (verbs V1 and V4 only appeared ungrammatically in S3). While participants were being trained on the language, presentation of one sentence from the ungrammatical chain was randomly interleaved within every three presentations of sentences from the grammatical chain. Subject-object noun pairs were randomized for each verb across presentations. There were a total of 96 training sentences.

Procedure. Participants underwent pre-training trials to acquaint them with the vocabulary. During pre-training they heard and saw each word along with pictures of each noun and scenes corresponding to each verb along with spoken audio of each noun/verb. All words were cycled through three times during pre-training. During the main experiment, all participants were told they were to learn an artificial language. They all saw a series of sentences describing computer animated scenes where a subject performed an action on an object. All sentences were presented in both spoken and written form.
Figure 3. Presentation of linguistic input in Experiment 1. The *generative learning* condition presented (a) positive examples generated by a speaker of the language and (b) negative examples generated by a non-speaker. The *discriminative learning* condition presented (c) positive and (d) negative examples as feedback to a prediction about grammaticality. Note that because verb-action pairings were randomized between subjects, the same verb does not correspond to the same actions in the different conditions.

In the *generative* condition, participants were told that they would listen to an adult speaker who was always spoke grammatical sentences and a child speaker who always spoke ungrammatically. Cartoon pictures of either the adult or child speaker accompanied each scene (see Figure 3 (a) and (b)). The child speaker's voice was low-pass filtered to create a believably child-like sound. We hypothesized that participants in this condition would behave
Similarly to a generative model: they would build a probabilistic representation of the language from the grammatical sentences produced by the adult speaker.

In the discriminative condition, participants were presented with spoken and written sentences describing each scene and asked to choose whether each of the presented sentences were grammatical or not. They were assured that only relevant words were used and they only had to figure out if the verb occurred in a grammatical location. Participants then received feedback on their choice (see Figure 3 (c) and (d)). For example, if a participant answered that the sentence was grammatical, they would see either “Yes, you were correct. This sentence is grammatical!” or “Sorry, you were incorrect. This sentence is ungrammatical!” The main difference from the generative condition is that in the discriminative condition, the presented sentences are assumed to be chosen at random, whereas in the generative learning condition, sentences from the adult speaker are assumed to have been sampled from the language distribution. We hypothesized that participants in the discriminative condition would behave similarly to a discriminative model: they would use feedback about both grammatical and ungrammatical sentences to formulate rules about what made sentences grammatical.

After the language learning phase, participants in both conditions were subjected to a grammar test. In this testing phase, participants were shown a series of written sentences and asked to rate the sentence as either grammatical or ungrammatical. Here, all sentences had blergen as the subject and nagid as the object. All verb-sentence structure combinations were shown twice. Additionally the verb V4 was shown an extra two times in S2 as this was the crucial generalization that we were testing.

Participants also underwent a production test in which they were shown a scene and asked to type in a sentence describing that scene. Because we did not want this to be a memory test, we displayed the relevant verb on the top of the screen. Pictures of all the
nouns, with their respective names below, were also available on the bottom of the screen for reference. Four scenes were presented for each verb, using subject-object noun pairs that were cycled through random. Verbs were also cycled through at random.

Results and Discussion

Figure 4 (a) shows the proportion of times each sentence was judged to be grammatical. The results suggest that participants in both conditions were largely able to learn much of the grammar structure. However, there were significant differences between the generative and discriminative conditions. Most notably, the generative learners overwhelmingly judged verb V4 to be ungrammatical in S2, while the majority of discriminative learners deemed V4 in to be grammatical in S2. This difference between

---

Figure 4. Results of Experiment 1. In this language, the absent construction was verb V4 in sentence structure S2. (a) Grammaticality judgments, showing proportion of times each sentence was judged to be grammatical. (b) Production results, showing proportion of productions made in each sentence structure for each verb. X denotes productions that were not in the form of any of the sentence structures.
conditions was highly statistically significant by a Pearson's $\chi^2$ test ($\chi^2 (1) = 7.28, p < .01$). This difference aligned with the difference in the predictions of the HBM (generative) model and the logistic regression (discriminative) model discussed earlier. Our results strongly suggest participants in the generative condition were learning language with a probabilistic perspective that allowed them to learn restrictions on verb alternations by using indirect negative evidence whereas participants in the discriminative condition made sampling assumptions that did not allow them to learn the alternation restriction.

Another difference we found between the two conditions was that discriminative learners were more willing to consider verbs to be alternating (i.e. allow those verbs to be grammatical in two sentence structures.) This is evidenced by the fact that participants in the generative condition rated occurrences of V1 (the alternating verb) in S1 and S2 as grammatical only 68% and 72% of the time. This is because many participants judged V1 to be grammatical in either S1 or S2 and not both. In fact, only 40% of participants in the generative condition learned the alternation for V1. On the other hand, participants in the discriminative condition rated occurrences of V1 in S1 and S2 grammatical 100% of the time (see Figure 3a). Pearson's $\chi^2$ tests for the number of participants who learned the V1 alternation correctly between conditions was significant, with $\chi^2 (1) = 4.62, p < .05$. From post-experiment questioning, we learned that many participants in the generative condition did not think verbs would occur in two possible sentence structures. None of the participants in the discriminative condition were constrained by this assumption.

Figure 4(b) shows the proportion of times each sentence was produced in the production test. These results showed that participants tended to use verbs in the sentences structure that they heard them in, with similar probabilities to the linguistic input. Notably, even though the majority of the learners in the discriminative condition rated verb V4 in S2 as grammatical, only 20% of the productions of V4 were in S2. This indicates that
participants in the discriminative condition did learn the distributional information in the language, but did not use it in the same way as the participants in the generative condition to assess grammaticality. This result is also in line with previous results that show that how often a sentence structure is produced is proportional to how often that structure is heard, and rarely heard structures are rarely produced, even if they are believed to be grammatical (Wonnacott et al., 2008).

**Experiment 2: An extended language**

The results of Experiment 1 suggested that people are sensitive to the sampling assumptions behind the linguistic input they receive, using indirect negative evidence only when the assumptions of the generative approach are warranted. However, one possible criticism of this experiment is that the artificial language we used had two verbs that unambiguously did not alternate, and one verb that unambiguously alternated. This could potentially create a bias against alternation, which might differ in salience in the generative and discriminative learning conditions. To rule out this alternative explanation, Experiment 2 used the same procedure to examine the effect of sampling assumptions on use of indirect negative evidence when the language contains a greater representation of alternating verbs. Participants were trained on an extended version of the language used in Experiment 1, doubling the number of alternating verbs. The language now contained two unambiguously alternating verbs and two unambiguously non-alternating verbs, removing any potential bias. The additional one-syllable verb was *glim*, corresponding to the transitive action *shrink-to-distance*.

**Method**

**Participants.** A total of 50 participants (25 in each condition) were recruited from the University of California, Berkeley community. Participants were compensated at a rate of $12/hour.
Stimuli. Stimuli were generated in the exact same fashion as in Experiment 1. The only difference is that we wished to further balance the ratio of alternating to non-alternating verbs. Thus, an additional alternating verb was introduced along with an additional transitive action, for a total of five verbs. Aside from the additional alternating verb (V2), the rest of the language was exactly the same as that used in Experiment 1, with V3 and V4 in the new language corresponding to V2 and V3 in the original language, and V5 being the new exception verb. Also, because now there were more verbs to learn, we increased the number of training trials so that participants viewed each verb 36 times (vs. 24 times in Experiment 1) for a total of 180 trials. The relative proportion of verbs shown in the various sentence structures and the semi-randomized method of generating a presentation order remained the same as in Experiment 1. Again, there is one verb (V5) that is never shown in one sentence structure (S2).

Procedure. The procedure was exactly the same as in Experiment 1, except for one alteration: Immediately before participants in both conditions began their training session, they were shown a screen that said, “Note: Some verbs may be grammatical in more than one context.” This was to encourage the participants to allow for the possibility of alternating verbs.

Results and Discussion

Figure 5(a) shows the proportion of times each sentence was judged to be grammatical. Our results suggest that participants in both conditions were largely able to learn much of the grammar structure, though with a little less accuracy than in Experiment 1, in which there was one fewer verb to learn. Again, there were significant differences between the generative and discriminative conditions, which aligned with the different predictions of the HBM (generative) model vs. the logistic regression (discriminative) model: The generative learners overwhelmingly judged the exception verb, V5, to be ungrammatical in
Figure 5. Results of Experiment 2. In this language, the absent construction was verb V5 in sentence structure S2. (a) Human grammar judgments, showing proportion of times each sentence was judged to be grammatical. (b) Human production results, showing proportion of productions made in each sentence structure for each verb. X denotes productions that were not in the form of any of the sentence structures.

S2, while the majority of discriminative learners allowed V5 to alternate, judging it as grammatical in S2. This difference between conditions was highly statistically significant by a Pearson's $\chi^2$ test ($\chi^2(1) = 7.611, p < 0.01$). Here, in comparison with Experiment 1, participants were more encouraged to allow for verb alternations: There were more alternating verbs in the language and the instructions explicitly said that alternations were possible. Still we find that participants in the generative condition restricted alternations on the exception verb when those in the discriminative condition did not.

As in Experiment 1, we found that discriminative learners were still more willing to consider verbs to be alternating (i.e. allow those verbs to be grammatical in two sentence structures). Participants in the generative condition rated S1 and S2 as grammatical 78% and
82% of the time for V1 (the first alternating verb) 76% and 76% of the time for V2 (the second alternating verb) in S1. This is compared with participants in the discriminative condition who rated S1 and S2 as grammatical 98% and 96% of the time for V1 and 100% and 100% of the time for V2 ($\chi^2(1) = 20.26, p < 0.00001$). Why the two conditions prompted significantly different prior assumptions about the prevalence of verb alternations will be a question for future research.

Figure 5(b) shows the proportion of times each sentence was produced in the production test. As in Experiment 1, participants in both conditions showed low production probability for the exception construction, showing that even in the discriminative condition, distributional information was learned but just not used as evidence for ungrammaticality.

**General Discussion**

We explored whether use of indirect negative evidence depends on the assumptions that learners make about how their linguistic input is generated. Our simulations showed that a generative learner will use indirect negative evidence, while a discriminative learner will not. The predictions of these simulations were borne out in two experiments with human learners. In our experiments, participants in two conditions viewed exactly the same sentences and were told whether each sentence was grammatical or ungrammatical. What varied between conditions was the way the grammaticality information was presented. In the discriminative condition, participants were given direct feedback on the grammaticality of sentences that were assumed to be sampled at random. Because of the random sampling assumption, the absence of a verb in a given sentence structure did not provide indirect negative evidence against the grammaticality of that construction, and people generalized accordingly. In contrast, participants in the generative condition were told that they were seeing sentences generated from the language, and used the absence of a particular construction as evidence against its grammaticality.
Our simulations and behavioral results begin to clarify the connection between formal analyses of language learning and human behavior. In particular, they indicate that adults are capable of adopting the assumptions of either the generative or the discriminative approach when learning an artificial language, modifying their inferences based to match the way in which the linguistic input is generated. This finding parallels previous work showing that adults and children are sensitive to sampling assumptions when learning novel words (Xu & Tenenbaum, 2007a, 2007b), suggesting that such a sensitivity might underlie language learning more generally.

The sensitivity to sampling assumptions demonstrated in our experiments has several implications for formal analyses of language learning. First, it suggests that human language learning might be best characterized by models that make sampling assumptions that match the generative process underlying the linguistic input people receive. To the extent that naturalistic language learning might seem better characterized by the assumptions of the generative approach – that people hear sentences sampled from the distribution associated with the language. This suggests that generative models may give a more accurate account of human language learning. Participants in our generative condition heard sentences spoken by a grammatical speaker, similar to the way children learn by listening to adult speech. In post-experiment questioning, many participants in this condition also stated that they ignored all negative evidence from the ungrammatical child speaker, similar to the way children often ignore negative evidence in real language acquisition. These observations support the idea that naturalistic human language learning is better characterized by the generative approach, complementing the recent success of generative models in predicting the results of other experiments (e.g., Perfors et al., 2010).

A second implication of our findings is that models of language learning should be capable of showing the same sensitivity to sampling assumptions as human learners. Since
most previous formal models of language learning choose to formulate the problem either in
terms of generative or discriminative learning, it is common to try to force the learning
situation to conform to the assumptions of the model. For example, a model that only
supports discriminative learning would treat the two conditions of our experiment in exactly
the same way, focusing just on the frequencies with which positive and negative examples of
each type were presented. An important message from our results is that these frequencies –
the kind of information summarized in Tables 1 and 2 – are not enough to fully describe the
situation involved in learning an artificial language. The assumptions about the way in which
the sentences are generated also play a critical role in the conclusions that people draw, and
need to be incorporated into any complete model of human learning.

The analyses presented in this paper suggest some interesting directions for future
research. First, any conclusions we draw about human language learning need to remain
tentative until it is clear that children display a similar sensitivity to sampling assumptions.
Testing whether child language learners recognize the difference between generative and
discriminative learning is thus an important direction for future research. It may be that
children can only adopt one of these approaches to language learning, which would
potentially provide further insight into the validity of these two ways of framing the problem
of language acquisition. Second, our experiments showed an interesting effect in which it
seemed that generative learners had a stronger expectation for verbs to be non-alternating.
Understanding the nature and origin of this bias, and how it relates to the assumptions of
generative learning, is interesting in itself, but perhaps even more interesting is the idea that
learners might adopt different inductive biases, and not just different sampling assumptions,
when using these two different approaches to language learning.

Our work presents a novel perspective of two main approaches to modeling language
acquisition, showing how these approaches differ in the use of indirect negative evidence.
We show that these two approaches correspond to two distinct ways of formulating the learning problem, which result in different sampling assumptions. Our experimental results show that human learners can learn artificial languages using both these approaches. These results pave the way for future research that can ascertain the contexts in which each approach is used in learning, providing insight into the factors that inform the remarkable human capacity for language.
Acknowledgements

Preliminary results of Experiment 1 were presented at the 22nd Neural Information Processing Systems conference. The authors thank Ky Merritt for assistance in data collection for the experiments. This work was supported by grant number RES-000-22-3275 from the Economics and Social Research Council and grant number SES-0631518 from the National Science Foundation.
Appendix: Model details

We illustrated the predictions of generative and discriminative models using logistic regression for the discriminative model and a hierarchical Bayesian model (HBM) for the generative model. This Appendix provides the details of the implementation of these two models.

Logistic regression

A discriminative model learns a mapping from sentences to grammaticality. If we observe \( n \) sentences, \( x_1 \ldots x_n \), each sentence \( x_i \) will be associated with a variable \( c_i \) indicating whether the sentence is grammatical (\( c_i = +1 \)) or ungrammatical (\( c_i = -1 \)). This information is represented by a feature vector \( f(x_i) \) that uses dummy variables to encode the verb, the sentence structure, and the interaction of the two (i.e. each sentence's particular verb and sentence structure combination). With \( m \) verbs and \( k \) sentence structures, this results in \( m \) verb features, \( k \) sentence structure features, and \( mk \) interaction features, each of which take the value 1 when they match the sentence and 0 when they do not. For example, a sentence containing the second of four verbs in the first of three sentence structures would be encoded with the binary feature vector 0100100001000000000, where the features are concatenated in the order (verb) (sentence structure) (interactions). The logistic regression model learns which features of sentences are predictive of grammaticality. This is done by defining the probability of grammaticality as

\[
p(c_i = +1 | x_i, \eta, b) = \frac{1}{1 + \exp\{-\eta^T f(x_i) - b\}}
\]

Where \( \eta \) and \( b \) are the parameters of the model. These parameters are estimated by maximizing the log likelihood \( \sum_{i=1}^{n} \log(p(c_i | x_i, \eta, b)) \). Features for which the likelihood is uninformative (e.g. features that are not observed) have weights that are set to zero.
Hierarchical Bayesian Model

A generative model captures the distribution of data using positive examples. Here we implement this using an HBM, also known to statisticians as a Dirichlet-Multinomial model (Gelman, Carlin, Stern, & Rubin, 2003; Kemp, Perfors, & Tenenbaum, 2007). With statistical notation an HBM can be written as:

\[ \alpha \sim \text{Exponential}(\lambda) \]
\[ \beta \sim \text{Dirichlet}(\mu) \]
\[ \theta \mid \alpha, \beta \sim \text{Dirichlet}(\alpha \beta) \]
\[ y^i \mid n^i, \theta \sim \text{Multinomial}(\theta) \]

where \( y^i \) is the data (i.e. the distribution of observed grammatical sentence structures) given \( n^i \) occurrences of verb \( i \), \( \theta \) is the distribution over sentence structures associated with verb \( i \), and \( \alpha, \beta \) describe higher-level generalizations about these distributions (see the main text for an intuitive description).

The input vector, \( y^i \) is a \( k \)-dimensional vector where each entry is the number of occurrences of verb \( i \) in the \( k \)th sentence structure. For example if V1 occurred in S1 2 times and S2 4 times and S3 0 times, \( y^1 \) would be [2 4 0] and \( n^1 \) would = 6. This model assumes fixed number of verbs and fixed number of possible sentence structures. \( \theta \) is a \( k \) dimensional vector of multinomial parameter values for the \( i \)th verb. For example, if there are two possible sentence structures, if the \( i \)th verb occurs 90% in the first sentence structure and 10% in the second, \( \theta \) would = [0.9 0.1]. This determines the probability of that \( i \)th verb occurs in the \( j \)th sentence structure, with observations of verb \( i \) in sentence structure \( j \) being drawn independently at random with probability \( \theta_j^i \). \( \theta \) is drawn from a Dirichlet distribution parameterized by \( \alpha \) and \( \beta \). The parameter \( \beta \) represents the distribution of sentence structures across all verbs while the parameter \( \alpha \) represents the extent to which each verb tends to appear in only in one sentence structure.
As described in the main text, this model makes inferences at Level 1 (values for $\theta$), and Level 2 (values for $\alpha$ and $\beta$), but Level 3 knowledge is fixed by setting $\lambda=1$ and $\mu=1$. This assumes weak prior knowledge that the range of $\alpha$ and $\beta$ do not contain extreme values. (Extensions to this model can also learn values of $\lambda$ and $\mu$.) The model is fit to the data by computing the posterior distribution $p(\Theta^i \mid y) = \int p(\Theta^i \mid \alpha, \beta, y) p(\alpha, \beta \mid y) d\alpha d\beta$. We estimate this distribution using a Markov Chain Monte Carlo (MCMC) scheme to perform numerical integration. Following Kemp et al. (2007), we use Gaussian proposals on $\log(\alpha)$, and proposals for $\beta$ are drawn from a Dirichlet distribution with the current $\beta$ as its mean. The simulations presented in the paper used 10,000 iterations of MCMC.
References


