Bayesian Learning at the Syntax-Semantics Interface

Sourabh Niyogi (niyogi@mit.edu)
Massachusetts Institute of Technology
Cambridge, MA USA

Abstract

Given a small number of examples of scene-utterance pairs of a novel verb, language learners can learn its syntactic and semantic features. Syntactic and semantic bootstrapping hypotheses both rely on cross-situational observation to hone in on the ambiguity present in a single observation. In this paper, we cast the distributional evidence from scenes and syntax in a unified Bayesian probablistic framework. Unlike previous approaches to modeling lexical acquisition, our framework uniquely: (1) models learning from only a small number of scene-utterance pairs (2) utilizes and integrates both syntax and semantic evidence, thus reconciling the apparent tension between syntactic and semantic bootststrapping approaches (3) robustly handles noise (4) makes prior and acquired knowledge distinctions explicit, through specification of the hypothesis space, prior and likelihood probability distributions.

Learning Word Syntax and Semantics

Given a small number of examples of scene-utterance pairs of a novel word, a child can determine both the range of syntactic constructions the novel word can appear in and inductively generalize to other scene instances likely to be covered by the concept represented (Pinker 1989). The inherent semantic, syntactic, and referential uncertainty in a single scene-utterance pair is well-established (c.f. Siskind 1996). In contrast, with multiple scene-utterance pairs, language learners can reduce the uncertainty of which semantic features and syntactic features are associated with a novel word.

Verbs exemplify the core problems of scene-utterance referential uncertainty. Verbs selectively participate in different alternation patterns, which are cues to their inherent semantic and syntactic features (Levin 1993). How are these features of words acquired, given only positive evidence of scene-utterance pairs?

The *syntactic bootstrapping* hypothesis (Gleitman 1990) is that learners exploit the distribution of "syntactic frames" to constrain possible semantic features of verbs. If a learner hears /glip/ in frames of the form /S glipped G with F/ and rarely hears /S glipped F into G/, the learner can with high confidence infer /glip/ to be in the same verb class as /fill/ and have the same sort of argument structure. A different distribution informs the learner of a different verb class. Considerable evidence has mounted in support of this hypothesis (c.f. Naigles 1990, Fisher et al 1994). In contrast, the *semantic bootstrapping* hypothesis (Pinker 1989) is that learners use what is common across scenes to constrain the possible word argument structures. If a learner sees a liquid undergoing a location change when /S glipped F/ is uttered, then /glip/ is likely to be in the same verb class as /pour/ and have the same sort of meaning.

Both hypotheses require the distribution of cross-situational observations. Prior accounts to model word learning have either ignored the essential role of syntax in word learning (Siskind 1996, Tenenbaum and Xu 2000), or require thousands of training observations (Regier et al 2001) to enable learning. In this paper we present a Bayesian model of learning the syntax and semantics of verbs that overcomes these barriers, by demonstrating how word-concept mappings can be achieved from very little evidence, where the evidence is information from both scenes and syntax.

Bayesian Learning of Features

We illustrate our approach with a Bayesian analysis of a single feature. On some accounts, verbs possess a *cause* feature which may be valued `1`, `*`, or `0` (Harley and Noyer 2000); depending on the value of the *cause* feature, the verb may appear in frame F1, F0, or both:

`1`	Externally caused - Ex: touch, load
	F1: He touched the glass.
	F0: *The glass touched.
`*`	Externally causable - Ex: break, fill
	F1: He broke the glass.
	F0: The glass broke.
`0`	Internally caused - Ex: laugh, glow
	F1: *He laughed the children.
	F0: The children laughed.

Assuming this analysis, learners who hear utterances containing a novel verb, not knowing the value of its *cause* feature, must choose between 3 distinct hypotheses H_1, H_*, and H_0. Clearly, one utterance cannot uniquely determine the value of the feature: if learners hear F1 (/S Ved O/), the feature sup-

ports H_1 or H_*; similarly, if learners hear F0 (/O Ved/), the feature may be H_0 or H_*. Two utterances cannot determine the feature uniquely either. Learners might receive both F1 and F0, supporting H_* uniquely. But they may also accidentally receive 2 utterances of the same form (F0, F0 or F1, F1), thus not resolving the ambiguity. If learners received 6 utterances of the same form F0 or F1, however, then there is overwhelming support for H_0 or H_1 respectively, and H_* seems far less likely.

A Bayesian analysis renders the above analysis precise and quantitative. Knowledge is encoded in three core components: (1) the structure of the hypothesis space $\mathcal{H}$; (2) the prior probability $p(H_i)$ on each hypothesis H_i in $\mathcal{H}$, before learners are provided any evidence; (3) the likelihood of observing evidence X given a particular H_i, $p(X|H_i)$. Given evidence $X = [x_1, \ldots, x_N]$ of N independent observations, by Bayes' rule the posterior probability of a particular hypothesis H_i is:

$$p(H_i|X) = \frac{\prod_{j=1}^{N} p(x_j|H_i)p(H_i)}{p(x_1, \ldots, x_N)} \quad (1)$$

signaling the support for a particular hypothesis H_i given evidence X.

In this case, x_j is the observation of a syntactic frame (F0 or F1), and X is a distribution of syntactic frames. One simple prior probability model $p(H_i)$ has each of the 3 hypotheses are equally likely, encoding that a verb is equally likely to be of the /touch/, /laugh/ or /break/ class:

$$p(H_1) = p(H_*) = p(H_0) = \frac{1}{3} \quad (2)$$

and a likelihood model $p(x_j|H_i)$ encoding how likely we are to observe frames F0 or F1 for the 3 different feature values of *cause*:

$$\begin{aligned} p(x_j = F1|H_1) &= .95 \quad p(x_j = F0|H_1) = .05 \\ p(x_j = F1|H_*) &= .50 \quad p(x_j = F0|H_*) = .50 \\ p(x_j = F1|H_0) &= .05 \quad p(x_j = F0|H_0) = .95 \end{aligned} \quad (3)$$

The above likelihood model says that when a verb has *cause*=1, we expect frames of the form /S Ved O/ 95% of the time; when a verb has *cause*=0, we expect /O Ved/ 95% of the time; when a verb has *cause*=*, we expect both syntactic frames.

Both the prior probability model and likelihood model are *stipulated*, encoding a learner's prior knowledge of grammar. Given these probability models, this allows for explicit computation of the support of each hypothesis. Suppose a learner receives F0. Then the support for each of the 3 hypotheses may be computed to be:

$$\begin{aligned} p(H_1|F0) &= \frac{(.05)(.33)}{(.05 + .50 + .95)(.33)} = .033 \\ p(H_*|F0) &= \frac{(.50)(.33)}{(.05 + .50 + .95)(.33)} = .333 \\ p(H_0|F0) &= \frac{(.95)(.33)}{(.05 + .50 + .95)(.33)} = \mathbf{.633} \end{aligned} \quad (4)$$

Any number of situations may be analyzed as such:

	Evidence X	$p(H_1\|X)$	$p(H_*\|X)$	$p(H_0\|X)$
1	F0	.033	.333	**.633**
2	F0, F0	.002	.216	**.781**
3	F0, F0, F0, F0, F0, F0	2e-8	.021	**.979**
4	F0, F1	.137	**.724**	.137
5	F0, F1, F0, F1, F0, F1	.007	**.986**	.007
6	F0, F1, F1, F1, F1, F1	**.712**	.288	5e-6

When only F0 is given as evidence (situation 1), while both H_0 and H_* are consistent with the observation, H_0 is nearly twice as likely. However, with 2 observations of F0 (situation 2) or 6 observations (situation 3), it is increasingly likely that H_0 is the correct hypothesis. With both F0 and F1 as evidence (situation 4), in contrast, H_* is far more likely; with more evidence (situation 5), it becomes more so. Finally, if the first frame is a "noise" frame and followed by 5 representative frames of F1 (situation 6), then H_1 is more likely instead.

Given this framework, just one or two observations is sufficient to make an informed judgement. Note that each additional observation increases certainty, and noise is handled gracefully.

Modeling Semantic Bootstrapping

In this section, we extend the single feature analysis to multiple features, where each feature represents information from scenes (from any modality, whether perceptual, mental, etc.). Setting aside verbal aspect, we may model possible verb meanings as a set of M features, where each feature represents a predicate on one or more of the arguments of the verb. For example, a set of single argument predicates might include:

moving(x), rotating(x), movingdown(x),
supported(x), liquid(x), container(x)

specifying the perceived situation about the argument of the verb (e.g. if it is moving, or moving in a particular manner, etc.) while a second set of two-argument predicates might specify the relationships between arguments, given that this is an externally caused (*cause*=1) event:

contact(x, y), support(x, y), attach(x, y)

Using these predicates, an idealized (partial) lexicon might contain the following word-concept mappings:

	cause	One arg x	Two arg x, y
/lower/	1	1*11**	11*
/raise/	1	1*01**	11*
/rise/	0	1*0***	
/fall/	0	1*1***	

specifying, in linear order, the value of each of the one and two-argument predicates above, e.g. that /lower/ has *cause*=1, moving(x)=1, rotate(x)=*, movingdown(x)=1, etc. – and thus its concept covers externally-cased motion events where an agent moves a theme downwards through supported contact. The verb /raise/ is nearly identical except it has movingdown(x)=0, while /fall/ and /rise/ involve internally-caused motion (*cause*=0) and do not

specify any two argument predicates. The values of * for the 4 rotating(x), liquid(x), container(x), and attach(x,y) predicates signal that these features are irrelevant to the verb's concept. Perception of a scene amounts to evaluating these predicates; scenes may or may not fall under the verb concept, conditioned on the values of these predicates. The presence of q of "irrelevant" features valued as * implies 2^q possible scenes consistent with the concept.

Given a hypothesis space of possible verb concepts formed by M of these sorts of predicates, the task of learning a verb's meaning given N observations $X = [x_1 \ldots x_N]$ of scenes, is to determine which of the 3^M possible concepts is the most likely. Just as before, a Bayesian model does so by computing the posterior probability distribution $p(H_i|X)$ over concepts, given a prior distribution on hypotheses $p(H_i)$ and a likelihood distribution of generating a particular x_j example given H_i:

$$p(x_j|H_i) = \begin{cases} \frac{1}{2^q} & \text{if } x_j \in H_i \\ 0 & \text{otherwise} \end{cases} ; p(H_i) = \frac{1}{3^M} \quad (5)$$

We can use Bayes' rule (Eq (1)) to compute the likelihood of any hypothesis given N independent examples. Intuitively, the above likelihood model says that out of the 2^q possible scenes that might fall under the concept H_i, all of them are equally likely; likewise, the prior probability model holds that all of the 3^M concepts are equally likely.

Consider a reduced hypothesis space where $M = 3$:

q	Concepts
0	`000, 001, 010, 011, 100, 101, 110, 111`
1	`00*, 01*, 10*, 11*, 0*0, 0*1, 1*0, 1*1, *00, *01, *10, *11`
2	`0**, *0*, **0, 1**, *1*, **1`
3	`***`

Given any distribution of scenes X, we can directly compute the posterior probability $p(H_i|X)$ of any of the 27 different concepts. Four are shown here, of increasing generality from a very specific concept (H_{000}) covering only one scene (`000`) to the most general concept H_{***} covering 2^M possible scenes:

	Observation X:	H_{000}	H_{00*}	H_{0**}	H_{***}
1	000	**.30**	.15	.07	.03
2	000, 000, 000	**.70**	.09	.01	.001
3	000, 001	.00	**.64**	.16	.04
4	000, 001, 000	.00	**.79**	.10	.01
5	000, 001, 000, 001, 000	.00	**.94**	.03	.001
6	000, 101, 010, 111, 000	.00	.00	.00	**1.0**

A single scene observation `000` is explained by all 4 hypotheses (situation 1) in a graded fashion. However, with 3 repeated observations (situation 2), most of the mass is concentrated on H_{000}. When scene observations require abstracting away irrelevant features, the more specific concepts must be discarded in favor of more general concepts (situation 3 vs 6). Each example consistent with the general concept further reduces ambiguity over the possible concepts (situation 4 vs 5).

Modeling Syntactic Bootstrapping

In this section, we demonstrate a Bayesian model of how the distribution of syntactic frames, as envisioned by Gleitman (1990), may be used to determine the semantic features of a verb. To do so, we introduce a new notion of *semantic agreement*, wherein features of a lexical head must *agree* with its complement. Consider the following idealized lexicon:

/fill/	fig: [0] con: [1]	/into/	fig: [1]
/pour/	fig: [1] liq: [1]	/with/	fig: [0]
/load/	fig: [*]	/glass/	con: [1]
		/water/	liq: [1]

A lexical head /fill/ *agrees* with a complement of /a glass with water/ but not with /water into a glass/, because the lexical head and its complement have a value 1 along the *fig* dimension. Likewise, a lexical head /pour/ agrees with a complement of /water into a glass/ but not /a glass with water/, because of the opposite value of *fig*. Finally, a lexical head such as /load/, because $*$ agrees with 0 and 1, accepts both complements. Thus, both /load the wagon with hay/ and /load hay into the wagon/ are valid derivations. A large number of verb classes can be seen to pattern into three classes along different feature dimensions in this way (Nomura et al 1994).

Any number of feature dimensions may be hypothesized, and may include selectional features, such as /fill/ requiring a container (con:[1]) or /pour/ requiring a liquid (liq:[1]) as its complement.

Suppose a learner hears /S glipped a glass with water/. The features of the novel verb /glip/ are unknown and the features of its complement /a glass with water/ are known. For the *fig* feature dimension of /glip/, there are 3 possible values, with 3 corresponding hypotheses H_0, H_1, H_*. As before, one observation is insufficient to infer H_0, as H_* is also possible. The following likelihood model for an unknown verb feature value V and the feature value of its complement C agreeing can be used for each feature dimension (fig, loc, con, etc.) to compute a probability distribution over the H_i:

$p(V, C)$	$V = 0$	$V = 1$	$V = *$
$C = 0$	.22	.01	.11
$C = 1$	.01	.22	.11
$C = -$	.11	.11	.12

Intuitively, the above says that with high probability, V and C agree, and with low probability (i.e. .01), they do not agree. The above joint distribution encodes both the prior distribution on V and the conditional distribution $p(C|V)$:

$$p(V = 0) = p(V = 1) = p(V = *) = \frac{1}{3} \quad (6)$$

$$\begin{array}{ll} p(C = V|V = 0 \text{ or } 1) = .65 & \\ p(C \neq V|V = 0 \text{ or } 1) = .03 & p(C = 0, 1|V = *) = .32 \\ p(C = *|V = 0 \text{ or } 1) = .32 & p(C = *|V = *) = .35 \end{array} \quad (7)$$

Given an assumption of perfect knowledge of the feature values of the complement, over multiple observations, the distributional evidence X in support of the 3 hypotheses can be readily evaluated. We can test how different distributions of syntactic frames correctly yield different probability distributions of a verbs syntactic and semantic features; this is thus a Bayesian model of Gleitman's (1990) "syntactic bootstrapping". Suppose a learner gets 4 syntactic frames of /glip/, all of the form /S glipped O with Z/. This is equivalent to having 4 perfect observations of fig:[0], which we annotate as $X = \texttt{0000}$. Then the likelihood $p(X|V)$ and posterior probability $p(V|X)$ of the 3 possible hypotheses can be evaluated directly via Bayes' rule:

Likelihood $p(X\|V)$	Posterior $p(V\|X)$
$p(X\|V=0) = (.65)^4$	$p(V=0\|X) = .941$
$p(X\|V=1) = (.03)^4$	$p(V=1\|X) = .000$
$p(X\|V=*) = (.32)^4$	$p(V=*\|X) = .059$

This is shown below, along with other distributions of syntactic frames:

Sit	Utterances (X)	$V=0$	$V=1$	$V=*$
1	4 /S Ved O with Z/ (0000)	**.941**	.000	.059
2	4 /S Ved O/ (****)	.292	.292	**.416**
3	2 /S Ved O with Z/, 2 /S Ved O into Z/ (0011)	.032	.032	**.936**
4	2 /S Ved O/, 2 /S Ved O with Z/ (**00)	**.769**	.000	.230
5	23 /S Ved O with Z/ 10 /S Ved O/	**1.00**	.000	.000
6	23 /S Ved O with Z/ 5 /S Ved O into Z/ 10 /S Ved O/	**.960**	.000	.040

With only 4 examples, the uncertainty of the value of the feature V is rapidly reduced (situations 1-4). As the number of examples increases (situation 4 vs 5), the evidence supports "all-or-none" or "rule-like" behavior, even with a significant number noisy frames (situation 5 vs 6).

Modeling Integrated Syntactic and Semantic Bootstrapping

We now integrate the two forms of bootstrapping described above, where given a distribution of both scenes and syntactic frames, a probability distribution over concepts consistent with both sources of evidence is determined. Consider the following possible syntactic frames:

Utterance	**u**	Attention
/Glipping!/	***	-
/S glipped water from a glass/	1**	W
/S glipped water into a glass/	1**	W
/S glipped water/	***	W
/S glipped a glass with water/	0**	G
/S glipped a glass/	***	G

and perceptually-derived semantic features of scenes:

Scene **s**	*Description/Semantic Features*
pour-fill	Person pouring water into a glass, filling it
G_{001}	Glass: Manner: None (0) State: Full (1)
W_{110}	Water: Manner: Pouring (1) State: None (0)
splash-fill	Person splashes water into a glass, filling it
G_{001}	Glass: Manner: None (0) State: Full (1)
W_{120}	Water: Manner: Splashing (2) State: None (0)
spray-fill	Person sprays water into a glass, filling it
G_{001}	Manner: None (0) State: Full (1)
W_{130}	Manner: Spraying (3) State: None (0)
pour-empty	Person pouring water out of glass, emptying it
G_{002}	Manner: None (0) State: Empty (2)
W_{110}	Manner: Pouring (1) State: None (0)
splash-empty	Person splashes water out of glass, emptying it
G_{002}	Manner: None (0) State: Empty (2)
W_{120}	Manner: Splashing (2) State: None (0)
pour-none	Person pouring some water into a glass
G_{000}	Manner: None (0) State: None (0)
W_{110}	Manner: Pouring (1) State: None (0)
spray-none	Person sprays water into a glass
G_{000}	Manner: None (0) State: None (0)
W_{130}	Manner: Spraying (3) State: None (0)

where features are ordered as:

fig, manner-of-motion, change-of-state

for each utterance **u** and scene possibility **s**. The subscripts on G and W annotate the observation of that argument for each of the 3 dimensions.

We may describe, just as before, how the cross-situational distributional evidence X of N independent scene-utterance pairs:

$$X = [(\mathbf{s}_1, \mathbf{u}_1), \ldots, (\mathbf{s}_N, \mathbf{u}_N)] \quad (8)$$

yields different word-concept mappings $p(H_i|X)$ through independent combination of the two sources of evidence:

$$p(H_i|X) = \frac{\prod_{j=1}^{N} p(\mathbf{s}_j|H_i)p(\mathbf{u}_j|H_i)p(H_i)}{p(X)} \quad (9)$$

For expository purposes, we will consider how the learner would rank each of the 6 precise hypotheses, and will assume they entertain only these:

English Verb	Hypothesis	Feature
pour	H_{pour}	11*
spray	H_{spray}	12*
splash	H_{splash}	13*
fill	H_{fill}	0*1
empty	H_{empty}	0*2
move	H_{move}	1**

The likelihood $p(\mathbf{s}_j|H_i)$ for each of the D independent dimensions ($D = 3$) is:

$$p(\mathbf{s}_j = s_1 \ldots s_D|H_i) = \prod_{k=1}^{D} p(s_k|H_i) \quad (10)$$

where our model for scene observations along the kth dimension is:

$$p(s_k|H_i) = \begin{cases} 1 - d_k\epsilon & \text{if} s_k = 0, H_i^k = * \\ \epsilon & \text{if} s_k \neq 0, H_i^k = * \\ 1 - d_k\delta & \text{if} s_k = H_i^k, H_i^k \neq * \\ \delta & \text{if} s_k \neq H_i^k, H_i^k \neq * \end{cases} \quad (11)$$

We annotate the value of the kth dimension of hypothesis H_i as H_i^k above. The first two lines model that when a feature is not valued ($H_i^k = *$), then scenes typically have 0 for the kth dimension ($d_1 = 2; d_2 = 3; d_3 = 3$), but do not match with probability ϵ. That is, observing pouring, spraying, splashing manners ($s_2 = 1, 2,$ or 3), and observing filling, emptying, or breaking change-of-states ($s_3 = 1, 2,$ or 3)

Situation	Scene **s**		Utterance **u**		H_{pour}	H_{spray}	H_{splash}	H_{fill}	H_{empty}	H_{move}
1	pour-fill	$\{G_{001}, W_{110}\}$	/S glipped water into a glass/	(1**)	**.889**	.008	.008	.000	.000	.093
2	pour-fill	$\{G_{001}, W_{110}\}$	/S glipped glass with water/	(0**)	.000	.000	.000	**.990**	.009	.000
3	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(***)	**.468**	.004	.004	**.468**	.004	.049
4	none		/S glipped water into a glass/	(1**)	.246	.246	.246	.004	.004	.254
5	none		/S glipped glass with water/	(0**)	.007	.007	.007	**.485**	**.485**	.007
6	none		/Glipping!/	(***)	.166	.166	.166	.166	.166	.170
7	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(***)						
	pour-empty	$\{G_{002}, W_{110}\}$	/S glipped water from a glass/	(1**)	**.998**	.000	.000	.000	.000	.001
	pour-none	$\{G_{000}, W_{110}\}$	/S glipped water/	(***)						
8	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(***)						
	splash-fill	$\{G_{001}, W_{120}\}$	/S glipped a glass with water/	(0**)	.000	.000	.000	**.999**	.000	.000
	spray-fill	$\{G_{001}, W_{100}\}$	/S glipped a glass/	(***)						
9	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(***)						
	splash-empty	$\{G_{001}, W_{120}\}$	/S glipped water/	(***)	.064	.064	.064	.000	.000	**.808**
	spray-none	$\{G_{001}, W_{100}\}$	/S glipped water/	(***)						

Figure 1: Word-concept mapping $p(H_i|X)$, given scene-utterance evidence X of a novel verb, /glip/

is far less likely than observing no manner of motion ($s_2 = 0$) or change of state ($s_3 = 0$) at all. Since observing a different value $s_j \neq 0$ is unlikely to have occurred by accident, it may be an important feature to the concept. The second two lines of (11) model that if a feature is valued ($H_i^k \neq *$), then scenes typically match that feature in value, but do not match with probability δ. That is, for example, given hypothesis H_{pour}, then most of the scenes will contain pouring in them. In our examples, $\epsilon = .1, \delta = .01$; qualitatively, results are not sensitive to changes in these values.

The output of our model is shown in Figure 1.

Suppose, as in Situation 1, a learner is given a single scene-utterance pair (pour-fill, /S glipped water into the glass/): $X = [(\mathbf{s}_1 = \{G_{110}, W_{110}\}, \mathbf{u}_1 = \mathtt{1}**, W)]$, and we wish to compute $p(H_i|X)$ for all $H_i \in \mathcal{H}$. We assume the learner can attend to the argument so as to extract relevant features from the scene. Given the scene pour-fill paired with utterance /S glipped water into a glass/, our Bayesian model places high weight on H_{pour}.

In Situation 2, the scene is the same, but now the syntax /S glipped a glass with water/ provides the learner with the information to attend not to the water's manner-of-motion but to the glass' change of state. Given $X = [(\mathbf{s}_1 = \{G_{110}, W_{110}\}, \mathbf{u}_1 = \mathtt{0}**, G)]$ our model weights H_{fill} heavily.

In Situation 3, the scene is the same, but now the syntax /Glipping!/ gives the learner less information, since the argument in the scene that the speaker may be referring to is unknown: $X = [(\mathbf{s}_1 = \{G_{110}, W_{110}\}, \mathbf{u}_1 = ---)]$ If there are A arguments in the scene, the speaker must have had a particular argument z in mind. The learner must *condition* on all the possibilities of z:

$$p(\mathbf{s}_j|H_i) = \sum_{a=1}^{A} p(\mathbf{s}_j|H_i, z_a)p(z_a) \qquad (12)$$

If learners consider all arguments equally salient ($p(z_i) = \frac{1}{A}$) then this effectively models /Glipping!/ as equivalent to /S is glipping Z1/ with probability $p(z_1) = .5$ and /S is glipping Z2/ with probability $p(z_2) = .5$. For simplicity, we assume $A = 2$ where Z1 is water, Z2 is the glass – but further referential uncertainty can be modeled with higher A. Because of the conditioning on each of A possibilities, this yields a less certain word-concept mapping.

In situation 4 through 6, the same syntactic frames are provided as in situations 1 through 3, but without the scene information. When some syntactic information is provided by the frame (situation 4, /S is glipping water into a glass/), then the manner-of-motion locative verbs are preferred over the change-of-state locative verbs, but no differentiation is possible without the scenes. Likewise, when the frame provides the opposite cue (situation 5, /S is glipping a glass with water/), the opposite preference is achieved, again with no differentiation between possible change-of-state verb concepts. When zero syntactic information is available (situation 6, /Glipping!/), all hypotheses prove equally likely.

Whereas in situation 3 the verb-concept mapping was ambiguous, primarily between H_{pour} and H_{fill}, in situation 7 and 8, learners are provided 2 additional examples to disambiguate. Both the scenes and syntactic frames in situation 7 support H_{pour}, while in situation 8 the scenes and syntactic frames support H_{fill}.

Finally, in situation 9, 2 different scene-utterance pairs primarily support the "superordinate" concept H_{move}, and not any "subordinate" manner-of-motion concept H_{pour}, H_{splash}, or H_{spray}.

Discussion

The reason why our analysis is able to infer so much from so little evidence is because so much is embedded in the given knowledge sources:

- the structure of the hypothesis space $\mathcal{H}$. Our examples contained a small number of feature dimensions and their possible values, but these may

be specified by interfaces to perceptual, motor, memory, or other "theory" representations. If so, whether these are innate or acquired are conditional on their source.

- priors $p(H_i)$ on hypotheses in $\mathcal{H}$. We used equal priors, but updating $p(H_i)$ based on language input is natural. In the verbal domain, such phemonena are commonly observed (e.g. manner vs. path, tight/loose-fit biases).
- likelihood of scenes $\mathbf{s}$ given the word concept $p(\mathbf{s}_j|H_i)$. We stipulated static values of ϵ and δ, but this can be acquired from observation.
- perfect knowledge of the features of the complement. We made this simplifying assumption to illustrate the essential elements of our model, but learners must acquire these features in parallel.
- the likelihood of agreement, $p(C|V)$, between a feature of a novel verb V and its complement C. We speculate that there is sufficient structure in partially learned words so as to acquire the structure in the joint distribution of feature values.

This richness of knowledge is in contrast to the models employed by Regier et al (2001) and Desai (2001), who train connectionist neural networks so as to learn the word-scene associations for adjectives/ nouns and verbs respectively. The high dimensionality of their models forces the need for thousands of training trials, and the interpretation of the weights is notoriously difficult. The assumptions behind these models are not justified by these authors. In contrast, our Bayesian approach makes the hypotheses, priors, and likelihoods explicit, holding this structure to be central.

Siskind (1996) views lexical acquisition as constraint satisfaction, and offers a robust algorithm where the mapping between input and hypothesis space is accomplished by pruning hypotheses that do not occur cross-situationally. Provided an idealized tokenization of the world, the algorithm does not need a large number of examples. However, Siskind's model does not yield any form of preference between different concepts, which is especially important when two or more concepts may be equally constrained by the data. We have shown how a Bayesian analysis explicitly yields preferences between concepts in the posterior probability distribution $p(H_i|X)$.

Tenenbaum and Xu (2000) take the important step of putting word learning in the Bayesian framework that we adopt here, showing how noun learning can occur with a small number of examples in a continuous-variable input space.

Crucially, however, the above models ignore the constraining role of syntax, despite considerable evidence that children use syntax to guide their verb-concept hypothesis space (Gleitman 1990, Naigles 1990, Naigles 1994, Fisher et al 1994, Snedeker and Gleitman 2002). Qualitatively, our models' performance matches the preferences of child learners, modeling their acquisition from as little as one example.

Our use of statistics does not imply any commitment to radical empiricism. Much prior knowledge is stipulated: the structure of the hypothesis space, the priors on hypotheses, and the likelihood of scene-utterance pairs given the hypotheses. It is not specified whether these stipulations are innate or themselves learnable. Linguistics and lexical semantics provide detailed theories of a much larger syntactic and semantic hypothesis space, and little prevents their inclusion in this framework.

Acknowledgements

Many thanks to Robert C. Berwick for motivating and supporting this work. Jesse Snedeker and Josh Tenenbaum provided many stimulating discussions. This work was funded by a provost grant to Prof. Joel Moses.

References

Desai, R. (2001). Bootstrapping in Miniature Language Acquisition. In *Proceedings of the Fourth International Conference on Cognitive Modeling*, pp. 61-66. Hillsdale, NJ: Erlbaum.

Harley, H. and Noyer, R. (2000) Licensing in the non-lexicalist lexicon. In Bert Peeters, (Ed.) *The Lexicon-Encyclopedia Interface*, Amsterdam: Elsevier Press.

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

Gleitman, L. (1990) The structural sources of verb meanings. *Language Acquisition*, 1990, 1:3-55.

Levin, B. (1993) *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago, IL.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, 117:357-374.

Nomura, N., Jones, D.A. and Berwick, R.C. (1994) An Architecture for a Universal Lexicon: A Case Study on Shared Syntactic Information in Japanese, Hindi, Bengali, Greek, and English. COLING 1994, 243-249.

Pinker, S. (1989) *Learnability and Cognition.* MIT Press, Cambridge, MA.

Regier et al (2001). The Emergence of Words. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.*

Siskind, J. (1996) A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings. *Cognition*, 61:39- 91

Snedeker, J. and Gleitman, L. (2002) Why it is hard to label our concepts. In G. Hall and S. Waxman (eds.), *Weaving a Lexicon*, Cambridge, MA: MIT Press.

Tenenbaum, J.B. and Xu, F. (2000) Word learning as Bayesian inference. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 517-522)