Bayesian Learning at the Syntax-Semantics Interface

Sourabh Niyogi niyogi@mit.edu
Massachusetts Institute of Technology
Cambridge, MA 02139, 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). Verbs which pattern together in the same constructions are commonly believed to have similar syntactic and semantic features. 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 /X glipped G with F/ and rarely hears /X 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 meaning. 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 posssible word argument structures. If a learner sees a liquid undergoing a location change when /X 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. We show what can be inferred from 1 example alone and how each additional example reduces the uncertainty of what the syntactic and semantic features can be.

Learning One Syntactic Feature

We illustrate our approach with a Bayesian analysis of a single feature. In the causative alternation there are two "frames" F0 and F1:

F0: Y Ved.
F1: X Ved Y.

Verbs possess a *cause* feature which may be valued 0, 1, or *: (Harley & Noyer 2000)

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_0, H_1, and H_*. Clearly, one utterance

cannot uniquely determine the value of the feature. If learners hear F1 (/X Ved Y/), the feature supports H_1 or H_*. Similarly, if learners hear F0 (/Y 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, then there is overwhelming support for H_0 or H_1 respectively.

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 $H_i \in \mathcal{H}$, without having any evidence, and (3) the likelihood of observing evidence X given a particular H_i, $p(X|H_i)$. Given evidence X of independent observations $x_1, \ldots, x_N$, 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 given evidence. In our case, x_j is the observation of a syntactic frame (F0 or F1), and X is a sequence of syntactic frames. For example, one prior probability model $p(H_i)$ has each of the 3 hypotheses are equally likely, encoding no information:

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

and a likelihood model $p(x_j|H_i)$ encoding a 5% error rate in observation of frames for the 3 different feature values:

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

Both of these are *stipulated*, encoding a learner's prior knowledge of the world. Given these probability models, this allows for *explicit* computation of the support of each hypothesis:

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

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.

Learning Multifeature Concepts

We now extend the single feature analysis to multiple features. Suppose learners encounter N utterances of /X is glipping Y/. From syntax alone, we can compute the probability that the *cause* feature is 0, 1, or *, but little else. However, when paired with scenes (from any modality, whether visual, auditory, emotional, and the like), additional information is available. The previous section showed how to learn the value of 1 feature, which happened to be a syntactic cause feature with 3 possible values. Multiple features can be combined together to form a larger hypothesis space. M 3-valued features constitute a 3^M size hypothesis space.

Setting aside verbal aspect, verb meanings as M features, each feature being some predicate on one or more of the essential arguments of the verb. Examples of possible predicates may be:

Cause(e)
One argument x: moving(x), rotate(x), movingdown(x), movingup(x), supported(x), liquid(x), container(x)
Two arguments x, y: contact(x, y), support(x, y), attach(x, y)

A verb like /lower/ may have most of the probability distribution weighted on $H_{1-1*101**-11*}$, /raise/ on $H_{1-1*101**-11*}$, /rise/ on $H_{0-1*01***}$, /fall/ on $H_{0-1*10**}$. Selectional predicates (e.g. liquid(x), container(x)) are no different than other predicates.

The task of learning a verb's meaning, given N observations $X = x_1, \ldots, x_N$ of scenes, is to determine the posterior probability distribution $p(H_i|X)$. Given a prior distribution on hypotheses $p(H_i^q)$ (q=No of 'dont-cares' in H_i) and a likelihood distribution of generating a particular $p(x_m)$ example given the hypothesis H_i^q:

$$p(H_i^q) = \frac{1}{3^M} \quad (4)$$

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

we can use Bayes' rule (Eq 1) to compute the likelihood of any hypothesis given N independent examples. Suppose we observe different sets of N samples, X, in a reduced hypothesis space where M=3. Then there are 27 possible concepts:

q	No	H_i Concepts
0	8	`000, 001, 010, 011, 100, 101, 110, 111`
1	12	`00*, 01*, 10*, 11*, 0*0, 0*1, 1*0, 1*1, *00, *01, *10, *11`
2	6	`0**, *0*, **0, 1**, *1*, **1`
3	1	`***`

Given scene observations X, we can directly compute the posterior probability $p(H_i|X)$ of any of the 27 different concepts, 4 shown here:

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

Just as in the one feature case, each example further reduces ambiguity over the possible concepts. Because knowledge is embedded in the prior and likelihood models, generalization is possible from very little evidence. This lower dimensionality is the source of inferential power.

Learning Lexeme Features

We now combine the two sources of information, syntactic and semantic, in a richer setting. Learning a novel word involves learning both types of features from utterance-scene pairs – syntactic features primarily from the utterance, semantic features primarily from the scene, but not exclusively. To illustrate our methodology, we will suppose a language learner has perfect knowledge of everything except the syntactic features and semantic features of a novel verb – e.g. word order, arguments, spatial and temporal reference – anticipating similar analyses can apply to handle these and other sources of uncertainty. We use a rudimentary form of minimalist operations (Chomsky 2000), although this is not essential to this approach.

A lexical item, possessing semantic and syntactic features, can merge with any other lexical item via a primitive operation Merge. When Merge events occur, the operation Agree (Chomsky 2000) checks features of both items to ensure that they match. A failure to match blocks the derivation, while a valid derivation must have all its features checked. Lexical items in a derivation in progress become inactive when their features are checked. For any given feature, there are 4 possible values:

`[0]` - valued, with value 0
`[1]` - valued, with value 1
`[*]` - unvalued, to be valued by derivation end
`[-]` - no value

with corresponding hypotheses H_0, H_1, H_*, H_-. For example, to derive /the glass/ the following two entries undergo Merge: (n=noun, d=determiner, *def*=definiteness)

/the/ n:[*] d:[1] def: [1]
/glass/ n:[1]

and an unvalued feature of /the/ becomes valued. Derivation /the the/ crashes due to an unvalued `n` feature. The result can merge with /see/ to generate /see the woman/.

Verbs which require a prepositional phrase (PP) argument have an unvalued feature, of the same type in common as the P. For example, /put/ has an unvalued loc feature (marked with *) which *must* get valued by /onto/:

/put/ cause:[1], dir:[1], loc:[*]
/onto/ loc:[1], contact:[1], dir:[1], fg:[1], ...

so as to generate /He put the books onto the shelf/ but not */He put the books/.

Verbs which optionally take a particular prepositional complement have a valued feature that agrees with the P. For example, /run/ has a valued *dir* feature which *can* agree with /into/:

/into/ dir: [1], b: [1], term: [1], loc: [1], fg: [1]
/run/ cause: [0], dir: [1]

so as to derive /He ran/ and /He ran into the room/. Likewise, the selectional criteria of load/pour/fill works through specification of the familiar "figure-ground" feature *fig* (1 if figure in Spec, 0 if ground in spec of P):

/load/ fig:[-]
/fill/ fig:[0] con:[1]
/pour/ fig:[1] liq:[1]

This *fig* feature is not specified for the verb class where both alternations are possible (e.g. load), but is valued 0 or 1 for the other 2 classes so as to agree with the preposition (/pour/ with locative prepositions such as /onto/, /cover/ with /with/). A considerable number of English verb alternation classes pattern in trios like this (see Levin 1993, Nomura et al 1994). We will assume that there is agreement between verb features and preposition features, where we assert one of many possible analyses: (loc=transfer of location, pos=transfer of possession, fig=figure in specifier of P, dir=dynamic, bnd= bounded, ter=terminal, con=container at terminal, cnt=contact)

	loc	pos	fig	dir	bnd	ter	con	cnt
/on/	1	-	1	-	-	-	-	1
/onto/	1	-	1	1	1	1	-	1
/off/	1	-	1	-	-	-	-	0
/in/	1	-	1	-	-	-	1	-
/into/	1	-	1	1	1	1	1	-
/with/	1	1	0	1	-	-	-	-
/from/	-	-	1	1	1	0	-	-
/of/	-	1	0	1	-	0	-	-
/to/	1	-	1	1	1	1	-	-
/toward/	1	-	1	1	0	1	-	1

The exact semantics of the features and their values are *not* relevant, only that they may be differentiated from each other.

Given this feature set, suppose a learner hears /He glipped the glass with water/ The probe is a novel verb /glip/ and the goal's features in /the glass with water/ are known. Because /water/'s features are not active, the relevant features are from /the glass/ and /with/:

/the glass/ d:[1] con: [1] n: [1]
/with/ n:[*] loc:[1] fig:[0] pos:[1] dir:[1]

Importantly, note that one observation is insufficient to infer whether /glip/ has a fig:[0] feature (or con:[1], and so forth), as it is also possible that /glip/ has fig:[*] or fig:[-], etc. With a likelihood model on agreement, for each feature dimension (fig, loc, con, etc.), a learner *can* compute a probability distribution of the four possible hypotheses. Given two items that Merge, an unknown probe verb P and a goal G, the joint probability distribution $p(P, G)$ for the 16 possibilities encodes knowledge that P and G must be in agreement with high probability:

p(P,G)	P=0	P=1	P=-	P=*
G=0	.165	.0025	.1225	.0825
G=1	.0025	.165	.1225	.0825
G=*	.0025	.0025	.0025	.0025
G=-	.08	.08	.0025	.08

Without such knowledge, inference is not possible.

The above distribution encodes both the prior distribution on P:

$$p(P = 0) = p(P = 1) = p(P = *) = p(P = -) = \frac{1}{4} \quad (6)$$

and the conditional distribution $p(P|G)$:

$$\begin{aligned} &p(P = G|G = 0, 1) = .443 \\ &p(P \neq G|G = 0, 1) = .007 \quad p(P = 0, 1|G = -) = .327 \\ &p(P = *|G = 0, 1) = .329 \quad p(P = *|G = -) = .001 \\ &p(P = -|G = 0, 1) = .221 \quad p(P = -|G = -) = .337 \end{aligned} \tag{7}$$

If we assume perfect knowledge of a feature of the Goal (i.e. the complement), then over multiple observations, the distributional evidence in support of the 4 Probe hypotheses (i.e. of the verb) can be readily integrated. Suppose a learner gets 4 utterance frames of /glip/, for example. If all 4 of the utterances are of the form /X Ved Y with Z/ then this is equivalent to having the Goal indicate 4 perfect observations of fig:[0], which we annotate as 0000. Then the likelihood $p(X|P)$ and posterior probability $p(P|X)$ of the 4 possible hypotheses can be evaluated directly via Bayes' rule:

Likelihood $p(X\|P)$	Posterior $p(P\|X)$
$p(X\|P = 0) = (.443)^4$	$p(P = 0\|X) = .732$
$p(X\|P = 1) = (.007)^4$	$p(P = 1\|X) = .000$
$p(X\|P = *) = (.329)^4$	$p(P = *\|X) = .222$
$p(X\|P = -) = (.221)^4$	$p(P = -\|X) = .046$

We can *test* how different distributions of syntactic frames correctly yield different probability distributions of syntactic features, amounts to Gleitman (1990) "syntactic bootstrapping" :

Utterances (X)	0	1	*	–
4 /X Ved Y with Z/ (0000)	.732	.000	.222	.046
4 /X Ved Y/ (----)	.319	.319	.000	.361
2 /X Ved Y with Z/, 2 /X Ved Y into Z/ (0011)	.001	.001	.828	.170
2 /X Ved Y/, 2 /X Ved Y with Z/ (--00)	.789	.000	.000	.210

As the number of examples increases, the evidence supports "all-or-none" or "rule-like" behavior, even in the presence of noisy frames:

Utterances (X)	0	1	*	–
23 /X Ved Y with Z/ (0) 1 /X Ved Y into Z/ (1) 1 /X Ved Y/ (-)	.998	.000	.002	.000
16 /X Ved Y with Z/ (0) 8 /X Ved Y into Z/ (1)	.000	.000	.998	.002
10 /X Ved Y with Z/ (0) 1 /X Ved Y into Z/ (1) 14 /X Ved Y/ (-)	.953	.000	.000	.047
1 /X Ved Y with Z/ (0) 1 /X Ved Y into Z/ (1) 23 /X Ved Y with Z/ (-)	.028	.028	.000	.944

The analysis above assumes perfect knowledge of the Goal, but we can relax this condition as well. If we do not have perfect knowledge of the Goal, and instead have only probability distributions on feature values of both Probe (verb) and Goal, then we can condition each computation on every probe-goal possibility given the current probability distribution, and improve the probability model for both, example by example.

As the above analysis demonstrates, it is possible to derive the semantics of one lexical item from the features of another item from utterances alone. However, utterance and scenes "agree" as well, and this regularity may be exploited.

Given N independent utterance-scene pairs X:

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

the two sources of evidence can be combined independently to compute $p(H_i|X)$:

$$p(H_i|X) = \frac{\prod_i p(\mathbf{s}_i|H_i)p(\mathbf{u}_i|H_i)p(H_i)}{p(X)} \tag{9}$$

In what follows, we will continue to assume that we have perfect information about word order, nouns, prepositions, and the like. This assumption about prior syntactic knowledge may be relaxed. Consider the following perceptually-derived semantic features:

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)

and possible syntactic "frames":

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

where features are ordered as:

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

For expository purposes we can consider how the learner would rank each of the 6 hypotheses, assuming they only entertain just the following:

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--

For each feature dimension, learners may have different priors on features having particular values. We consider by way of example a prior $p(s_j)$ structured as:

$$p(s_j = 0) = (1 - d_j\delta), p(s_j \neq 0) = \delta \tag{10}$$

where for small δ, the prior holds that usually, scenes have 0 for the jth dimension ($d_1 = 3, d_2 = 4, d_3 = 4$). Observing pouring, spraying, splashing manners ($s_2 = 1, 2, \text{or} 3$), and observing filling, emptying, or breaking change-of-states ($s_3 = 1, 2, \text{or} 3$) is far less likely than observing no manner of motion ($s_2 = 0$) or change of state ($s_3 = 0$) at all. Since observing

Situation	Evidence X: Scene S		Utterance U		Verb-Concept Mapping $p(H_i\|X)$: H_{pour}	H_{spray}	H_{splash}	H_{fill}	H_{empty}	H_{move}
1	pour-fill	$\{G_{001}, W_{110}\}$	/X glipped water into a glass/	(1–)	**.880**	.010	.010	.000	.000	.101
2	pour-fill	$\{G_{001}, W_{110}\}$	/X glipped glass with water/	(0–)	.000	.000	.000	**.989**	.011	.0001
3	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(—)	**.463**	.006	.006	**.463**	.005	.058
4	none		/X glipped water into a glass/	(1–)	.246	.246	.246	.004	.004	**.254**
5	none		/X 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	$\{G_{002}, W_{110}\}$	/X glipped water from the glass/	(1–)	**.998**	.000	.000	.000	.000	.002
	pour-none	$\{G_{000}, W_{110}\}$	/X glipped water/	(—)						
8	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(—)						
	splash-fill	$\{G_{001}, W_{120}\}$	/X glipped a glass with water/	(0–)	.000	.000	.000	**.999**	.000	.000
	spray-fill	$\{G_{001}, W_{100}\}$	/X glipped a glass/	(—)						
9	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(—)						
	splash-empty	$\{G_{001}, W_{120}\}$	/X glipped water/	(—)	.061	.066	.066	.000	.000	**.806**
	spray-none	$\{G_{001}, W_{100}\}$	/X glipped water/	(—)						

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

a different value $s_j \neq 0$ is unlikely to have occurred by accident, it may be an important feature to the concept.

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

$$p(\mathbf{s} = s_1 \dots s_D|H_i) = \prod_{j=1}^{D} p(s_j|H_i) \qquad (11)$$

where we allow each feature of s to be in the concept with probability $1 - \epsilon$ and inconsistent with ϵ:

$$p(s_j|H_i) = \begin{cases} 1-\epsilon & \text{if } s_j = H_{ij},\ H_{ij} \in \{0, \dots\} \\ \epsilon p(s_j) & \text{if } s_j \neq H_{ij},\ H_{ij} \in \{0, \dots\} \\ p(s_j) & \text{if } H_{ij} \in \{*, -\} \end{cases}$$

Conceptually, this says if we knew with absolute certainty that the hypothesis was in fact H_i (e.g. H_{pour}), then most of the scenes $\mathbf{s}$ we observe will contain pouring in them; if they don't contain pouring in them, then the kind of scenes that will be observed instead will be "generic". In our examples, $\epsilon = .1$; qualitatively, results are not aren't sensitive to small changes in epsilon.

Suppose, as in Situation 1 of Figure 1, a learner is given a single scene-utterance pair (pour-fill, /X glipped the water into the glass/): $X = (s_1 = \{G_{110}, W_{110}\}, u_1 = 1 - -, W)$, 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 (if not, it is as if no scene information is available). So for a particular hypothesis $H_{pour} = H_{11-}$:

$$\begin{aligned} p(X|H_{11-}) &= p(s_1 = \{G_{110}, W_{110}\}|H_{11-})p(u_1 = 1 - -|H_{11-}) \\ &= p(s = W_{110}|H_{11-})p(\text{P=1}|\text{G=1})p(\text{P=-}|\text{G=1})p(\text{P=-}|\text{G=-}) \\ &= (.9)^2(.7)(.443)(.327)(.337) \end{aligned}$$

The likelihood function $p(X|H_i)$ can be computed for each of the 6 hypotheses, weighted by $p(H_i)$, and normalized to compute the posterior probability $p(H_i|X)$:

H_i	Likelihood $p(X\|H_i)$	$p(H_i\|X)$
H_{pour}	(.9)(.9)(.7)(.443)(.327)(.337)	.880
H_{spray}	(.9)(.1.1)(.7)(.443)(.327)(.337)	.001
H_{splash}	(.9)(.1.1)(.7)(.443)(.327)(.327)	.001
H_{fill}	(.1.1)(.1)(.7.1)(.007)(.337)(.327)	.000
H_{empty}	(.1.1)(.1)(.7.1)(.007)(.337)(.327)	.000
H_{move}	(.9)(.1)(.7)(.443)(.337)(.337)	.101

This is also shown in figure 1. As expected, given the scene *pour-fill* paired with utterance /X glipped the water into the glass/, a learner with the above model should rationally conclude that the most likely hypothesis is in fact H_{pour}.

In Situation 2, the scene is the same, but now the syntax /X glipped the 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 $X = [s = \{G_{110}\}, u_1 = 0 - -, G])$ a similar computation yields that the most likely hypothesis is H_{fill}, shown in figure 1.

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: $X = (s_1 = (G_{110}, W_{110}), u_1 = - - -)$ 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:

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

If learners consider all arguments equally salient ($p(z_i) = \frac{1}{A}$) then this effectively models /Glipping!/ as equivalent to /X is glipping Z1/ with probability $p(z_1) = .5$ and /X 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.

$p(s = \{G_{001}, W_{110}\}|H_i) = .5p(s = W_{110}|H_i) + .5p(s = G_{001}|H_i)$

yielding different likelihood and posterior estimates, shown in figure 1.

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, /X 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, /X is glipping a glass with water/), the opposite preference is achieved, again with no differentiation be-

tween possible change-of-state verb concepts. When absolutely no 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 support a "superordinate" concept H_{move}, but 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 $p(H_i)$ (e.g. Manner vs. Path, tight/loose-fit) based on language input is natural.
- priors $p(s_j)$ on scenes having feature values. We stipulated static values of ϵ, but this can be acquired from observation.
- likelihood of scenes $\mathbf{s}$ given the word concept $p(\mathbf{s}|H_i)$. Again, this was stipulated, but could be acquired.
- perfect knowledge of the features of the argument (Goal) G. We made this simplifying assumption to illustrate the essential elements of our model, but learners must acquire these features in parallel.
- likelihood of agreement, $p(P|G)$, between a feature of a novel word P and its argument G. 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 model 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 make 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 algorithm does not contain any form of preference between hypotheses. In contrast, our form of analyses embeds preference information explicitly in the prior $p(H_i)$ and likelihood $p(X|H_i)$ functions.

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, each of 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.

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 nothing 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

Chomsky, N. (2001) Derivation by Phase. In *Ken Hale: A Life in Language*, Cambridge, MA: MIT Press.

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* (pp. 349-374), Amsterdam: Elsevier Press.

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

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.

Jackendoff, R. S. (1990) *Semantic Structures*. MIT Press, Cambridge, MA.

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 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)