Learning an Input Filter for Argument Structure Acquisition

Laurel Perkins
Dept. of Linguistics
University of Maryland
College Park, MD 20742
`perkinsl@umd.edu`

Naomi H. Feldman
Dept. of Linguistics & UMIACS
University of Maryland
College Park, MD 20742
`nhf@umd.edu`

Jeffrey Lidz
Dept. of Linguistics
University of Maryland
College Park, MD 20742
`jlidz@umd.edu`

Abstract

How do children learn a verb's argument structure when their input contains non-basic clauses that obscure verb transitivity? Here we present a new model that infers verb transitivity by learning to filter out non-basic clauses that were likely parsed in error. In simulations with child-directed speech, we show that this model accurately categorizes the majority of 50 frequent transitive, intransitive and alternating verbs, and jointly learns appropriate parameters for filtering parsing errors. Our model is thus able to filter out problematic data for verb learning without knowing in advance which data need to be filtered.

1 Introduction

Young language learners are limited by partial knowledge in identifying the structure of sentences they hear and of their language in general. This partial knowledge may lead to inaccurate parses of their input, resulting in data that are misleading about the true structure of their language. Here we investigate a problem of misleading data in verb learning: how learners identify verbs' syntactic properties despite the presence of unknown grammatical structures that obscure those properties.

We propose a new model for the acquisition of argument structure, the syntactic property of a verb that determines which types of clauses it can occur in (Chomsky, 1965; Chomsky, 1981; Grimshaw, 1990). We model how a learner can use verb distributions to infer whether a verb can occur in a transitive clause with both a subject and an object, an intransitive clause with only a subject, or both. This inference depends on the ability to accurately perceive the arguments in a clause: whether a clause has a subject and an object, or only a sub-

ject. Identifying these arguments is straightforward for "basic" clause types like (1) and (2), but more difficult for "non-basic" clause types that do not follow the subject-verb-object word order typical of English, like (3):

(1) John ate a sandwich. Amy threw a frisbee.

(2) John ate. (*Amy threw.)

(3) What did John eat? What did Amy throw?

A learner tracking when direct objects are present after the verb would notice that sentences like (1) contain both subjects and objects, and sentences like (2) contain only subjects. It would then follow that *throw* is obligatorily transitive whereas *eat* can alternate. But this strategy is complicated by the wh-object questions in (3). These questions do not have direct objects after the verb, but do have two arguments: the wh-word *what* stands in for the verb's object. These data may be misleading for a child who has not yet learned how to identify wh-questions in her language. She might note the absence of a direct object after the verb and perceive the sentences in (3) as intransitive, mistakenly concluding that *throw* can alternate just like *eat*.

To learn verb transitivity, learners need some way to filter out the non-basic clauses in their input; this ability is assumed in prominent theories of verb learning such as syntactic and semantic bootstrapping (Gleitman, 1990; Lidz and Gleitman, 2004; Pinker, 1984; Pinker, 1989). In this paper we present a Bayesian model that learns the parameters for such a filter solely by tracking the distributions of verbs with and without direct objects. The model does so under the assumption that some sentence observations will be generated in error, due to mis-parses of non-basic clauses like (3). In simulations with child-directed speech, we show that this model can learn the parameters for

filtering these parsing errors in order to categorize 50 verbs as transitive, intransitive, or alternating. We thus demonstrate that it is possible for a learner to learn an input filter for verb learning without knowing in advance what needs to be filtered.

2 Filtering Input

Many have proposed that learners need some way to filter the data they use in acquiring their language (Adriaans and Swingley, 2012; Lidz and Gleitman, 2004; Pearl and Lidz, 2009; Pinker, 1984). This filtering is important for theories of verb learning under which children rely on systematic relations between verbs' syntactic properties and their meanings, e.g. semantic and syntactic bootstrapping (Fisher et al., 2010; Gleitman, 1990; Landau and Gleitman, 1985; Lasnik, 1989; Pinker, 1984; Pinker, 1989). Non-basic clauses obscure these relations, so learners need a way to filter them out of the data they use for verb learning.

Pinker (1984) posits two solutions: either parents might do the filtering and avoid producing these sentences in their children's presence, or children might internally filter these sentences themselves. Parental filtering does not seem to occur: even before their second birthday, English-learning children hear a large number of wh-questions (between 10-17% of their total input), the majority of which do not follow the typical word order of English (Stromswold, 1995). These non-basic clause types are thus prevalent in child-directed speech, necessitating a different filtering solution.

The second logical solution is internal filtering: perhaps children can filter out non-basic clauses themselves. This proposal implicitly assumes that children know which sentences to filter out. However, experimental evidence suggests that children may not have the ability to recognize which sentences contain wh-questions before the age of 20 months, an age at which substantial verb learning is already taking place (Gagliardi et al., 2016). Furthermore, this ability may depend on prior verb knowledge. Identifying that sentences like (3) contain object wh-questions requires the learner to detect when a fronted phrase (like *what*, *who*, or *which NP*) stands in relation to a verb that needs a patient argument and is locally missing one. But this requires the learner to know which verbs take patient arguments, in order to notice when those arguments are needed and missing. In other words, the learner needs to detect that these sentences con-

tain direct object gaps, rather than intransitive uses of these verbs— but in order to do so, the learner must know which verbs are transitive.

The filtering problem thus risks being circular: learners need to know which verbs are transitive in order to detect the signals of non-basic clause types like object wh-questions, but they also need to filter out sentences containing these clause types in order to learn which verbs are transitive. Pinker (1984) posits that children might avoid this problem by using sentence meaning, context, and intonation to identify a filter on their input. Our approach, by contrast, does not require learners to know the parameters of the input filter before they are able to learn verbs. Instead of fixing one of these pieces of knowledge to learn the other, children may jointly infer which verbs are transitive and the parameters for filtering sentences containing non-basic clauses. We thus model a learner who can filter out errors in parsing non-basic clauses, without needing to first identify where those errors came from.
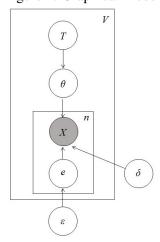
3 Model

Our model uses the distribution of direct objects within and across verbs to infer both verb transitivity and the parameters for filtering non-basic clauses. We adopt a Bayesian framework, in which a learner observes a data pattern and infers the probability of some properties of the system that may have generated that data. This framework conveniently allows us to specify the alternative systems (verb transitivity properties vs. mis-parses of non-basic clauses) that our learner considers for the verb distributions it observes. Our model follows other Bayesian approaches to argument structure acquisition (Alishahi and Stevenson, 2008; Perfors et al., 2010), but considers a different problem than the one explored in that literature. Instead of learning which verb classes exist in a particular language, our model is designed to solve the problem of learning which verbs map to which known transitivity classes despite input that obscures these mappings.

3.1 Generative Model

The model learns from observations of direct objects or no direct objects in sentences containing particular verbs. These observations are formalized as the Bernoulli random variable X in the graphical model in Figure 1. Each $X^{(v)}$ represents an observation from a sentence containing verb v in the model's input, with a value of 1 if the sentence

Figure 1: Graphical Model

contains a direct object and 0 if it does not. These observations can be generated by two processes: the transitivity of verb v, represented by the variables T and θ in the upper half of the model, or an internal parsing error, represented by the variables e, ϵ, and δ in the lower half of the model. We will describe each of these processes in turn.

In the upper half of the model, each $X^{(v)}$ is conditioned on the parameter $\theta^{(v)}$, a continuous random variable defined for values from 0 to 1 inclusive. This parameter controls how frequently a verb v will be used with a direct object: the learner assumes that for every observation $X^{(v)}$, a biased coin is flipped to determine whether the sentence contains a direct object, with probability $\theta^{(v)}$, or does not, with probability $1 - \theta^{(v)}$.

The parameter $\theta^{(v)}$ is conditioned on the variable $T^{(v)}$, which represents the transitivity of verb v. T is a discrete random variable that can take on three values, corresponding to transitive, intransitive, and alternating verbs. Each of these values determines a different distribution over θ. For the transitive category of T, θ always equals 1: the verb should always occur with a direct object. For the intransitive category, θ always equals 0: the verb should never occur with a direct object. For the alternating category, θ takes a value between 0 and 1 inclusive. The prior probability distribution over θ in this case is a uniform $Beta(1,1)$ distribution.

In the lower half of the model, each X is conditioned on a Bernoulli random variable e, which represents the input filter. If $e_i^{(v)} = 0$, the observation in $X_i^{(v)}$ was generated by $\theta^{(v)}$ and $T^{(v)}$, and accurately reflects the transitivity of verb v. But if $e_i^{(v)} = 1$, the observation in $X_i^{(v)}$ was generated by

an internal parsing error, meaning the learner did not have adequate grammatical knowledge to parse the sentence correctly. This observation was not generated by $\theta^{(v)}$ and $T^{(v)}$, and may not accurately reflect the transitivity of verb v, so it should be ignored for the purpose of inferring $T^{(v)}$. Each $e^{(v)}$ is conditioned on the variable ϵ, which represents the probability of an internal parsing error occurring for any verb in the input. The model learns a single parameter value for ϵ across all verbs.

The second parameter of the input filter is δ, which represents the probability of observing a direct object when an observation was generated by an internal parsing error. Thus, whether a sentence contains a direct object or no direct object depends on one of two biased coins. If $e_i^{(v)} = 0$ and the observation accurately reflects the verb's transitivity properties, then one biased coin is flipped and the sentence contains a direct object with probability $\theta^{(v)}$. If $e_i^{(v)} = 0$ and the observation was generated by a parsing error, then a different biased coin is flipped and the sentence contains a direct object with probability δ. Like ϵ, δ is a shared parameter across all verbs. We assume that both ϵ and δ have a uniform $Beta(1,1)$ prior distribution.

3.2 Joint Inference

We use Gibbs sampling (Geman and Geman, 1984) to jointly infer T, ϵ, and δ, integrating over θ and summing over e, with Metropolis-Hastings (Hastings, 1970) proposals for ϵ and δ.

We begin by randomly initializing ϵ and δ, and sampling values of T for each verb given values for those input filter parameters. From observations of a verb with and without direct objects, the model determines which value of T was most likely to have generated those observations. For $k^{(v)}$ direct objects in $n^{(v)}$ sentences containing verb v, we use Bayes' rule to compute the posterior probability of each value for $T^{(v)}$,

$$P(T^{(v)}|k^{(v)}, \epsilon, \delta) =$$
$$\frac{P(k^{(v)}|T^{(v)}, \epsilon, \delta) P(T^{(v)})}{\sum_{T^{(v)}} P(k^{(v)}|T^{(v)}, \epsilon, \delta) P(T^{(v)})} \quad (4)$$

Bayes' Rule tells us that the posterior probability of a particular value of T given $k^{(v)}$ and the other model parameters is proportional to the likelihood, the probability of $k^{(v)}$ given that value of T and those parameters, and the prior, the probability of

13

T before seeing any data. We assume that T is independent of ϵ and δ, and that all three values of T have equal prior probability.

To calculate the likelihood, we must sum over e. This sum is intractable, but because all of the values of e for the same verb and the same direct object status are exchangeable, we make the computation more tractable by simply considering how *many* errors were generated for sentences with and without direct objects for a particular verb. We divide the $k^{(v)}$ observed direct objects for a verb into $k_1^{(v)}$ direct objects that were observed accurately and $k_0^{(v)}$ direct objects that were observed in error. The total $n^{(v)}$ observations for verb v are likewise divided into $n_1^{(v)}$ accurate observations and $n_0^{(v)}$ errorful observations. We then calculate the likelihood by marginalizing over $n_1^{(v)}$ and $k_1^{(v)}$, again assuming independence among T, ϵ, and δ,

$$
p(k^{(v)}|T^{(v)}, \epsilon, \delta) =
$$

$$
\sum_{n_1^{(v)}=0}^{n^{(v)}} \left[\sum_{k_1^{(v)}=0}^{k^{(v)}} p(k^{(v)}|k_1^{(v)}, n_1^{(v)}, \delta) \right.
$$
$$
\left. p(k_1^{(v)}|n_1^{(v)}, T^{(v)}) \right] p(n_1^{(v)}|\epsilon) \quad (5)
$$

The first term in the inner sum is equivalent to $p(k_0^{(v)}|n_0^{(v)}, \delta)$, assuming we know $n^{(v)}$, the total number of observations for a particular verb. This is the probability of observing $k_0^{(v)}$ errorful direct objects out of $n_0^{(v)}$ errorful observations, which follows a binomial distribution with parameter δ. The second term in the inner sum is the probability of observing $k_1^{(v)}$ accurate direct objects out of $n_1^{(v)}$ accurate observations, which follows a binomial distribution with parameter $\theta^{(v)}$. Recall that $\theta^{(v)} = 1$ for the transitive category of T, and $\theta^{(v)} = 0$ for the intransitive category of T. For the alternating verb category, $\theta^{(v)}$ is unknown, so we integrate over all possible values of $\theta^{(v)}$ to obtain $\frac{1}{n_1^{(v)}+1}$. The last term in (5) is the probability of observing $n_1^{(v)}$ accurate observations out of the total $n^{(v)}$ observations for verb v, which follows a binomial distribution with parameter $1 - \epsilon$.

After sampling values for T for each verb in the dataset, we then sample values for ϵ and δ. If T denotes the set of values $T^{(1)}, T^{(2)}, ..., T^{(V)}$, and k denotes the full set of observations of direct objects $k^{(1)}, k^{(2)}, ..., k^{(V)}$ for all V verbs in the input, we can define functions proportional to the posterior

distributions on ϵ and δ, $f(\epsilon) \propto p(\epsilon|T, k, \delta)$ and $g(\delta) \propto p(\delta|T, k, \epsilon)$, as

$$
f(\epsilon) = p(k|T, \epsilon, \delta)p(\epsilon) \quad (6)
$$

$$
g(\delta) = p(k|T, \epsilon, \delta)p(\delta) \quad (7)
$$

where the likelihood $p(k|T, \epsilon, \delta)$ is the product over all verbs v of $p(k^{(v)}|T^{(v)}, \epsilon, \delta)$, as calculated in (5).

Within the Gibbs sampler, we resample ϵ using 10 iterations of a Metropolis-Hastings algorithm. We begin by randomly initializing ϵ. At each iteration, we propose a new value ϵ', sampled from the proposal distribution $Q(\epsilon'|\epsilon) = N(\epsilon, 0.25)$. Because the proposal distribution is symmetric, this new value is accepted with probability

$$
A = min\left(\frac{f(\epsilon')}{f(\epsilon)}, 1\right) \quad (8)
$$

If the new value ϵ' has higher probability given T, k and δ under equation (6), it is accepted. If it has lower probability under equation (6), it is accepted at a rate corresponding to the ratio of its probability and the probability of the old value of ϵ.

After sampling ϵ, we resample δ with 10 iterations of Metropolis-Hastings. The proposal and acceptance functions are analogous to those for ϵ.

We ran multiple chains from different starting points to test convergence of T, ϵ, and δ. For the simulations reported below, we ran 1,000 iterations of Gibbs sampling. We took every tenth value from the last 500 iterations as samples from the posterior distribution over T, ϵ, and δ.

4 Data

We tested the model on a dataset selected from the CHILDES Treebank (Pearl and Sprouse, 2013). We used four corpora of child-directed speech (803,188 total words), which were parsed using the Charniak or Stanford parser and hand-checked by undergraduates. See Table 1 for corpus details.

Our dataset consists of sentences containing the 50 most frequent action verbs in these corpora that could be characterized as transitive, intransitive, or alternating. We excluded verbs that were obligatorily ditransitive or frequently took clausal or verbal complements: these included mental state verbs (e.g. *want*), aspectual verbs (e.g. *start*), modals (e.g. *should*), auxiliaries (e.g. *have*), and light verbs (e.g. *take*).

We sorted these 50 verbs into transitive, intransitive, and alternating categories according to the

Corpus	# Children	Ages	# Words	# Utterances
Brown- Adam, Eve, & Sarah, (Brown, 1973)	3	1;6-5;1	391,848	87,473
Soderstrom (Soderstrom et al., 2008)	2	0;6-1;0	90,608	24,130
Suppes (Suppes, 1974)	1	1;11-3;11	197,620	35,904
Valian (Valian, 1991)	21	1;9-2;8	123,112	25,551

Table 1: Corpora of Child-Directed Speech.

English verb classes described in Levin (1993), supplemented by our own intuitions for verbs not represented in that work. These classes provide a target for learning meant to align with adult speaker intuitions, independent of the corpus data that the model learns from. The transitive and intransitive categories are conservative; verbs like *jump* are considered alternating even though they occur infrequently in their possible transitive uses (e.g. *jump the horses over the fence*). These target categories thus set a high bar for our model to reach.

We then conducted an automated search over the Treebank trees for the total occurrences of each verb in the corpora, in all inflections, and the total occurrences with overt (pronounced) direct objects. These direct object counts included transitive basic clauses like those in (1), but not wh-object questions with object gaps like those in (3). Table 2 lists these 50 verbs along with their counts and percentage occurrences with direct objects.

Verb	Total	% DO
Transitive Verbs		
feed	222	0.93
fix	338	0.91
bring	606	0.89
throw	313	0.88
hit	215	0.87
buy	362	0.84
catch	185	0.76
hold	580	0.70
wear	477	0.60
Alternating Verbs		
pick	332	0.90
drop	169	0.88
lose	187	0.86
close	166	0.85
touch	183	0.84
leave	356	0.83
wash	196	0.82
pull	331	0.81
open	342	0.77
push	357	0.77
cut	263	0.75
turn	485	0.72
build	299	0.72
bite	197	0.72
knock	161	0.71
read	509	0.69

Verb	Total	% DO
Alternating Verbs, cont.		
break	553	0.63
drink	369	0.61
eat	1324	0.59
sing	307	0.52
blow	256	0.52
draw	376	0.51
move	241	0.46
ride	284	0.35
hang	153	0.35
stick	194	0.29
write	583	0.27
fit	229	0.21
play	1571	0.20
stand	301	0.07
run	230	0.06
swim	184	0.05
walk	253	0.04
jump	197	0.04
sit	862	0.01
Intransitive Verbs		
wait	383	0.15
work	256	0.04
cry	263	0.03
sleep	454	0.03
stay	308	0.01
fall	611	0.00

Table 2: Counts and percentage uses with direct objects (DO) of 50 verbs in dataset.

5 Simulations

We tested our model on the dataset described in the previous section. We compare our model's performance to an oracle model that already knows the parameters of the input filter, and two baselines.

5.1 Joint Inference Model

Figure 2 displays the posterior probability distributions over verb categories inferred by our joint inference model for each verb. Black bars represent the probability assigned to the transitive category, dark gray bars represent the probability assigned to the intransitive category, and light gray bars represent the probability assigned to the alternating category. The true categories for each verb are shown below the horizontal axis. Figure 3 displays the posterior distributions inferred for ϵ and δ.

The percentage of verbs categorized correctly by the model is reported in Table 3. The model achieves highest accuracy in categorizing the intransitive verbs: for all but one of these verbs, the model assigns highest probability to the intransitive category. The exception is the verb *wait*, which the model assigns highest probability under the alternating category. This is due to prevalent uses of temporal adjuncts, as in *wait a minute*, that were parsed as direct objects in the CHILDES Treebank. Thus, a learner who likewise misparses these adjuncts as direct objects would infer that *wait* is an alternating rather than intransitive verb.

The model assigns 6 out of the 9 transitive verbs highest probability under the transitive category. Three transitive verbs are assigned highest probability under the alternating rather than the transitive category: *catch, hold,* and *wear*. This is likely because these verbs display different behavior than the other transitive verbs in the corpus. The verb *hold* occurs frequently in verb-particle constructions (e.g. *hold on*), which might be treated differently than simple verbs by learners. The verbs *catch* and *wear* appear to occur at much higher rates than other transitive verbs in non-basic clauses: *catch* occurs frequently in passive constructions (e.g. *get caught*), and *wear* occurs frequently in

Figure 2: Posterior Distributions over Verb Categories (T), Joint Inference Model
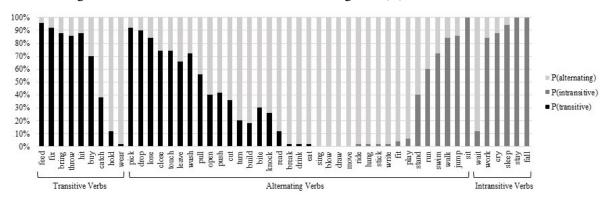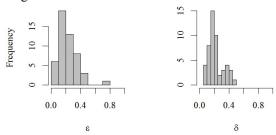

Figure 3: Posterior Distributions over ϵ and δ

wh-object constructions (e.g. *what are you wearing?*). Whether learners likewise mis-classify these verbs, or can accommodate different rates of missing direct objects within the set of transitive verbs, is a topic for further investigation.

The model assigns highest probability for most of the alternating verbs to the alternating verb category. There are 13 exceptions. The verbs *pick, drop, lose, close, touch, leave, wash*, and *pull* are assigned highest probability under the transitive category because they infrequently occur in their possible intransitive uses in child-directed speech. The verbs *run, swim, walk, jump,* and *sit* are assigned highest probability under the intransitive category because these verbs very infrequently occur in their possible transitive uses. Thus, the model over-regularizes the alternating verbs that alternate infrequently, preferring the more deterministic transitive and intransitive verb categories.

In order to evaluate the model's inference of ϵ and δ, we estimated the true values of these parameters in our dataset. The proportion of transitive verbs with missing direct objects in the dataset gives us an estimate of $(1 - \delta) \times \epsilon$, and the proportion of intransitive verbs with spurious direct objects (e.g. *wait a minute*) gives us an estimate

of $\delta \times \epsilon$. Solving these two equations, we find that $\delta = 0.18$ and $\epsilon = 0.24$. The posterior probability distribution over δ inferred by our model has a mean of 0.23, and the probability distribution over ϵ has a mean of 0.22. Our model thus slightly over-estimates the value of δ and under-estimates the value of ϵ. However, it infers values for these parameters close to the true values in the corpus, enabling it to infer the correct transitivity properties for 2/3 of the verbs in our dataset.

5.2 Oracle Model

To evaluate our model's performance, we compare it against an oracle model in which δ is fixed to 0.18 and ϵ to 0.24 in order to reflect their estimated true values in our dataset. This allows us to see how our model compares to a model that knows the parameters for the input filter in advance.

The posterior probability distributions over verb categories inferred by the oracle model are displayed in Figure 4. Our joint inference model performs identically to the oracle model with intransitive verbs, and almost as well with transitive verbs: the oracle model succeeds in identifying one more transitive verb, *catch*, as transitive. Our joint inference model performs better than the oracle model in categorizing alternating verbs: the oracle model has an even higher tendency to over-regularize the verbs that alternate infrequently.

Inferring the parameters of the input filter thus results in comparable, and maybe slightly better, accuracy in categorizing verbs than knowing these parameters in advance. It should be noted that the values of these parameters are important: when we run a version of the oracle model with inappropriate values for ϵ and δ, performance decreases substantially. Thus, our model performs comparably to

Model	% Transitive	% Intransitive	% Alternating	% Total Verbs
Joint Inference	0.67	0.83	0.63	0.66
Oracle	0.77	0.83	0.54	0.62
No-Filter Baseline	0.00	0.00	1.00	0.70
Random Baseline	0.33	0.33	0.34	0.34

Table 3: Percentages of Verbs Categorized Correctly.

Figure 4: Posterior Distributions over Verb Categories (T), Oracle Model

5.3 No-Filter Baseline

We've seen that our model accurately categorizes 2/3 of the verbs in our dataset by inferring appropriate parameters of a filter on its input, and performs comparably overall with a model that knows those parameters in advance. To determine how much the input filter matters in this inference, we compare our model to a baseline that lacks this filter. We can instantiate a model with no filter by setting ϵ to zero, representing zero probability of parsing errors. Because every verb in our dataset occurs some but not all of the time with direct objects, and this model assumes there are no parsing errors to filter out, it assigns every verb to the alternating category. It thus categorizes 100% of the alternating verbs correctly, achieving 70% overall accuracy because alternating verbs make up 70% of our dataset. However, this accuracy comes at the cost of failing to categorize any verbs as transitive or intransitive. Our joint inference model performs substantially better in this regard, categorizing the majority of transitive and intransitive verbs correctly. Thus, an input filter is important for differentiating alternating from non-alternating verbs.

5.4 Random Baseline

We finally compare our model against a baseline that assigns verbs randomly to transitivity categories, assuming that each value of T has equal prior probability. For each verb in our dataset this model flips a fair 3-sided coin to determine its transitivity category. This model thus categorizes 1/3 of the transitive, intransitive, and alternating verbs correctly, resulting in 34% overall accuracy. Our joint inference model performs significantly better on each verb class, and nearly twice as well overall.

5.5 Summary

We find that inferring an appropriate filter on the input matters for verb transitivity learning, but that the parameters of this input filter can be learned. Our model performs comparably to an oracle model that knows these values in advance. It performs substantially better in categorizing transitive and intransitive verbs than a baseline model that lacks an input filter altogether, and performs twice as well overall as a random baseline. These results demonstrate that our model is able to infer reasonable values for the input filter parameters, allowing it to accurately categorize the majority of transitive, intransitive, and alternating verbs.

6 Discussion

In this paper we introduce a model that infers the parameters of a filter on its input for argument struc-

ture acquisition. Our model accurately categorizes 2/3 of the most frequent transitive, intransitive, and alternating verbs in child-directed speech on the basis of their distributions with and without direct objects, by learning to filter out sentences that were likely mis-parsed. This enables the learner to avoid drawing faulty inferences about verb transitivity from non-basic clause types, such as wh-object questions, that may be mistaken for intransitive clauses. Our model performs substantially better than baseline models that lack an input filter and performs comparably to an oracle model that knows these input filter parameters in advance, demonstrating that this input filter both matters for verb learning and can be learned.

Our model offers a novel solution to the problem of identifying an appropriate input filter for verb learning (Lidz and Gleitman, 2004; Pinker, 1984) : where previous approaches have implicitly assumed that children must have a way of identifying the sentences to be filtered, our model learns an input filter without knowing its parameters in advance. Instead, the learner infers the input filter parameters jointly with verb transitivity. This reduces the prior knowledge needed for initial verb learning: the child does not need to identify which sentences likely contain wh-questions and other non-basic clause types as a prerequisite for learning which verbs are transitive. Note that we do not claim that the Bayesian joint inference performed by our model represents the exact algorithms performed by child learners; although there is substantial literature on young children's statistical inference capabilities (Gomez and Gerken, 2000), this model is intended only as a proof of concept that such joint inference is possible.

The model makes two types of errors in inferring verb categories. First, it is unable to correctly categorize some transitive and intransitive verbs that behave differently than other verbs in their category, such as *catch, hold, wear*, and *wait*. Further investigation is necessary to determine whether these verbs pose difficulties for child learners as well. A second type of error is over-regularizing alternating verbs that alternate infrequently: the model prefers to assign these verbs to the transitive and intransitive categories. This is an example of a learner preferring a more deterministic analysis for probabilistic input, a tendency also found in child learners in artificial language studies (Hudson Kam and Newport, 2009). The error-filtering mechanism

we present here could thus potentially provide a way to model other forms of over-regularization in learning.

Other future directions include extending this model to cross-linguistic data, particularly to languages with free object-drop. Chinese, Korean, and Japanese have a syntactic mechanism for dropping the direct object of any transitive verb, unlike in English where object-drop is a lexical property of specific verbs. As as result, learners of these languages might be subject to even higher rates of parsing errors if they perceive object-drop sentences as intransitive. For this reason, these languages are potentially problematic for syntactic bootstrapping strategies that rely on learners accurately identifying transitive verbs (Lee and Naigles, 2005; Lee and Naigles, 2008). But if appropriate parameters for filtering out problematic object-drop constructions in these languages can be inferred, our model may help address concerns with the feasibility of syntactic bootstrapping in these languages.

Finally, this model learns verb transitivity by effectively filtering out sentences containing non-basic clause types, without identifying what these clause types are. But children do eventually learn to identify non-basic constructions in their language. If children initially filter out certain sentences as parsing errors for the purposes of verb learning, they must eventually learn that many of these sentences are generated by systematic syntactic operations, such as those that create wh-questions in English. In future work, we aim to investigate whether a learner can identify which non-basic constructions are present in sentences that were parsed in error. Learning verb transitivity can likely help the child identify these constructions: if a child expects a direct object for a verb and encounters sentences where this object does not appear, that child may be compelled to examine those anomalous sentences to determine the cause of the missing object. Thus, a strategy of filtering out non-basic constructions by initially treating them as parsing errors may eventually help the learner identify not only verb transitivity, but also the nature of those non-basic constructions themselves.

References

Frans Adriaans and Daniel Swingley. 2012. Distributional learning of vowel categories is supported by prosody in infant-directed speech. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 72–77.

Afra Alishahi and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science*, 32:789834.

Roger Brown. 1973. *A First Language: The Early Stages*. Harvard University Press, Cambridge, MA.

Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Noam Chomsky. 1981. *Lectures on government and binding*. Foris Publications.

Cynthia Fisher, Yael Gertner, Rose M. Scott, and Sylvia Yuan. 2010. Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2):143–149.

Annie Gagliardi, Tara M. Mease, and Jeffrey Lidz. 2016. Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15-and 20-month-olds. *Language Acquisition*, 23(3):1–27.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE-PAMI*, 6:721–741.

Lila R. Gleitman. 1990. The structural sources of verb meanings. *Language acquisition*, 1(1):3–55.

Rebecca L. Gomez and LouAnn Gerken. 2000. Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4(5):178–186.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press, Cambridge, MA.

W. K. Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Carla L. Hudson Kam and Elissa L. Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59:30–66.

Barbara Landau and Lila R. Gleitman. 1985. *Language and Experience: Evidence from the Blind Child*. Harvard University Press.

Howard Lasnik. 1989. On certain substitutes for negative data. In Robert J. Matthews and William Demopoulos, editors, *Learnability and linguistic theory*, pages 89–105. Kluwer, Dordrecht.

Joanne N. Lee and Letitia R. Naigles. 2005. The input to verb learning in mandarin chinese: A role for syntactic bootstrapping. *Developmental Psychology*, 41(3):529–540.

Joanne N. Lee and Letitia R. Naigles. 2008. Mandarin learners use syntactic bootstrapping in verb acquisition. *Cognition*, 106(2):1028–1037.

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

Jeffrey Lidz and Lila R. Gleitman. 2004. Yes, we still need Universal Grammar. *Cognition*, 94:85–93.

Lisa Pearl and Jeffrey Lidz. 2009. When domain-general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development*, 5:235–265.

Lisa Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68.

Amy Perfors, Joshua B. Tenenbaum, and Elizabeth Wonnacott. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37:607–642.

Steven Pinker. 1984. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Steven Pinker. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.

Melanie Soderstrom, Megan Blossom, Irena Foygel, and James L. Morgan. 2008. Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, 35:869–902.

Karen Stromswold. 1995. The Acquisition of Subject and Object Wh-questions. *Language Acquisition*, 4(1/2):5–48.

Patrick Suppes. 1974. The semantics of children's language. *American Psychologist*, 29:103–114.

Virginia Valian. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition*, 40(1-2):21–81.