Pragmatics-utilizing distributional learner (PUDL) without deterministic hypothesis space

Boram Kim
Department of Linguistics
University of California, Los Angeles
bkim21@ucla.edu

Joonsuk Kang
Department of Statistics
University of Chicago
joonsukkang@uchicago.edu

Abstract

We introduce Pragmatics-Utilizing Distributional Learner (PUDL) to simulate verb transitivity learning in 15-month-old English learners. The model incorporates pragmatic reasoning about question-answer relations in neutral wh-questions. Our proposal outlines a developmental trajectory that features a temporary overregularization stage where learners generalize all verbs into one category, due to difficulty in distinguishing Prepositional Phrases from Noun Phrase objects. The results demonstrate the effectiveness of a pure distributional model enhanced by pragmatic knowledge in addressing learning challenges posed by noisy input.

1 Introduction

Learning how verbs behave in terms of taking direct objects proves to be a challenging task for learners. The complexity of verb transitivity learning arises from messy data that learners encounter, as illustrated in (1-2).

(1) *transitive*

a. Alex threw the truck.

b. What did Alex throw?

c. *Alex threw.

(2) *intransitive*

a. I waited.

b. I waited for Alex.

In an ideal setting for learning verb transitivity, learners would be exposed solely to examples featuring transitive verbs with a direct object (1a) and intransitive verbs without any adverbials (2a). With this transparent input, they could seamlessly make inferences about the transitivity patterns of verbs. However, reality often deviates from this ideal, exposing learners to less-than-optimal examples. In the surface string of (1b), the transitive verb 'throw' does not take any direct object due to the English rule of question formation: a direct object moves to the beginning in object wh-questions. A learner who hasn't yet acquired this non-local wh-dependency might be misled to infer from (1b) that 'throw' does not always take a direct object. In contrast, 'wait' is an intransitive verb that does not take a direct object, as demonstrated by the contrast between (1c) and (2a). Confusingly for learners though, prepositional phrases (PP) such as 'for Alex' in (2b) often occur with the intransitive verb 'wait.' Novice learners, who have yet to acquire the distinction between the PP 'for Alex' in (2b) and the NP 'the truck' in (1a), might incorrectly infer from utterances like (2b) that 'wait' is a transitive verb. The abundance of utterances like (1b) and (2b) in learners' input prompts a critical question: How do learners, faced with such misleading input, eventually arrive at accurate generalizations that transitive verbs like 'throw' consistently require a direct object, while intransitive verbs like 'wait' do not take a direct object?

Assuming grammatical knowledge of learners around 15 months old, we hypothesize that (i) pragmatic reasoning is what enables them to realize questions like (1b) do not serve as evidence for the intransitive nature of transitive verbs, but (ii) due to the failure to distinguish NP arguments (e.g., 'the truck' in (1a)) from PP adjuncts (e.g., 'for Alex' in (2b)) at the proposed developmental stage, they undergo a temporary overregularization stage where they perceive all verbs as transitives, on their way to the final destination, i.e., adult grammar. Our assumption about the developmental timeline is directly motivated by experimental results. Behavioral studies show that 15-month-olds behave as if they comprehend wh-dependency in (1b) (Gagliardi et al., 2016; Perkins and Lidz, 2021). [1] On the other hand, it has been experimen-

[1] In this regard, our claim about pragmatic reasoning can

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 86-95.
Irvine, California, June 27-29, 2024

tally shown that children as young as 19 months old incorrectly interpret PPs (she's wiping *with the tig*) as denoting a patient, a thematic role typically expressed by a direct object (she's wiping *the tig*) (Lidz et al., 2017). In other words, the learners we assume have overcome the learning problem that arises in the transitive domain (1), but not in the intransitive domain (2).

To model our target learner, English-learning 15-month-olds capable of pragmatic reasoning, but not PP vs. NP resolution, we propose Pragmatics-Utilizing Distributional Learner (PUDL). Using Bayesian Information Criterion (BIC), we show that the PUDL goes through an overregulazation phase where it prefers all verbs to be transitives. Compared to pure distributional learner (DL), which is not pragmatically informed, the PUDL's performance is farther from true knowledge about verb transitivity patterns, when asked to cluster verbs into three groups (transitive, alternating, intransitive). Still, pragmatic reasoning is hypothesized to be crucial to grappling with the misleading data of (1b) kind in the transitive domain; once learners become question-savvy, they are not tricked anymore by (1b). The resulting overregularization inference that every verb takes a direct object is inevitable given the messy nature of data they receive in the intransitive domain; 15-month-olds frequently hear utterances like (2b), while perceiving PPs incorrectly as NP objects.

The proposal is consistent with the idea that regularization, in general, plays a pivotal role in both first and second language acquisition (e.g., Hudson Kam and Newport 2005; Austin et al. 2022). Furthermore, we show that a pure distributional learner, as opposed to a learner with additional inductive bias, such as filtering (Perkins et al., 2022), is just as promising to tackle the puzzle in verb transitivity learning, although a full comparison with the PUDL augmented by the PP vs. NP resolution is left for future research.

2 Pragmatics-utilizing distributional learner (PUDL /pudəl/)

We propose Pragmatics-Utilizing Distributional Learner (PUDL), a pure distributional model that sidesteps deterministic hypotheses as part of its inductive bias but is bolstered by pragmatic knowledge. The base model we start with is distributional learner (DL). In the proposed model, verb categories do not have a fixed direct object (DO) probability; instead, they have probability distributions over the interval [0,1]. Intuitively, our learner operates with confidence in the received data, compared to alternative learners that filter out some proportion of data for successful learning. Without knowing that the input is noisy, the PUDL perceives every piece of data, including (1b) and (2b), as a valuable signal, as is reasonable to be assumed for learners as young as 15 months old who have no clue about deterministic verb transitivity. We assume that all they are sensitive to is the distributional patterns of verb transitivity.

The central challenge for our base model concerns a transition to acquiring correct deterministic knowledge without relying on a predefined deterministic hypothesis space. We propose that pragmatic understanding of discourse context plays a crucial role in addressing this issue for transitive verbs. Specifically, recognizing that (1b) functions as a neutral question that seeks information facilitates learners' transitivity acquisition. For instance, let's assume, for illustrative purposes, that the verb 'throw' occurs in the form of (1a) 80% of the time in the input, while 20 % of the time, it takes the form of (1b). Based on the observations from the input, a learner would form immature knowledge that 'throw' occurs with a direct object only 80% of the time. Once pragmatically informed, however, the learner associates the remaining 20% or so, due to (1b), with the information-seeking discourse function inherent in wh-questions. It would then cease its search for a missing direct object in interrogative sentences, recognizing that such information-seeking sentences are supposed to lack a direct object, i.e., the *answer* of the question. This nuanced yet straightforward pragmatic reasoning prompts the learner to update the initial underestimated knowledge about 'throw.' As a result, the learner moves closer to the correct understanding that 'throw' should always occur with a direct object, ideally reaching near 100%. The gap, previously attributed to 'throw's intrinsic property, is now ascribed to a specific discourse context of information seeking, which allows verbs to lack a direct object.

Two concerns may arise regarding (i) whether the complexity of the proposed pragmatic reasoning is appropriate for a learner as young as 15 months old, and (ii) imperfect correlation between missing direct objects and questions. First, despite the dis-

be understood as an attempt to answer how such knowledge of wh-dependency arises in young learners.

course function of questions being more complex than its declarative counterpart, two factors are hypothesized to enhance learners' capacity for the proposed pragmatic reasoning: (i-a) the prevalence of questions in child-directed speech, verifiable from corpus, and (i-b) distinctive rising intonation associated with questions. On the second point (ii) regarding the imperfect correlation, it is true that not every noise in the data takes the form of question. For example, transitive verbs used in relative clauses (3a) and in passives (3b) also lack direct objects. The noisy input of these kinds would prevent even the question-savvy learner from reaching 100%, i.e., acquiring deterministic knowledge found in adult grammar.

(3) a. I found the truck Alex threw.

b. The truck was thrown.

We assume that a learner at this stage, where they just start to distinguish questions from non-questions, indeed fails to attain 100% correct knowledge about verbs' transitivity property. Understanding complex constructions such as relative clauses and passives likely happens later in a child's life, whether it involves a pragmatic process or not. The upshot is that the presence of other kinds of misleading data such as (3) does not argue against the plausibility of the PUDL's learning schema and the proposed developmental trajectory.

A more serious challenge to the PUDL is that not all questions take the exact form of object wh-question in (1b). Polar questions (4a), rising declaratives (4b), and subject wh-questions (4c) do not lack direct objects even though they are questions.

(4) a. Did you throw the truck?

b. You threw the truck?

c. Who threw the truck?

However, polar questions (4a) and rising declaratives (4b) involve different discourse contexts from those of wh-questions in that they are *biased*. Buring and Gunlogson (2000) argue that positive polar questions like (4a) are not neutral; they can be felicitously asked in the presence of compelling contextual evidence. Similarly, rising declaratives, extensively studied in semantics, are biased questions, where the addressee might be asked for information, but the speaker is not neutral in their expectation (see, for example, Farkas and Roelofsen (2017) for formal modeling of the latter discourse behavior). Therefore, it is reasonable to assume that a learner can distinguish the discourse function of neutral wh-questions (seeking information without any expectations; (1b)) from non-neutral polar questions or rising declaratives (4a-b), which express the speaker's bias or may not necessarily expect an information-bearing answer.

Furthermore, the questions in (4) do not pose a challenge for verb transitivity learning in the first place. While a learner at the proposed stage may not correctly parse or understand each question in (4), the data are not misleading in terms of learning verb transitivity because 'throw' has a direct object in all three questions of (4). We proceed with the assumption that subject wh-questions of the (4c) kind are not noisy and, therefore, do not influence the learner's transitivity acquisition during the assumed developmental phase. In this phase, the transitivity-learning learner grapples with transparently noisy data, such as the example given in (1b). Whenever a violation of transitivity is observed as in (1b) (modulo relative clauses and passives), the PUDL associates the utterance with its unique discourse context, that is, seeking information by asking a question, and treats it as occuring with a direct object, even if the utterance (1b) lacks a direct object on the surface.

3 Data

The data we utilized are several corpora of child-directed speech from CHILDES (MacWhinney, 2000), specifically Brown (Brown, 1973), Soderstrom (Soderstrom et al., 2008), Suppes (Suppes, 1974), and Valian (Valian, 1991). Regarding the selection of corpora and the specific set of verbs, we followed Perkins et al. (2022) for a transparent comparison (Section 6). To model verb transitivity learning, they chose the 50 most frequent action verbs, classified into transitive, alternating, and intransitive categories.

Given our goal to model a learner around 15 months old, who has not yet resolved the NP vs. PP distinction, our learner blindly treats many elements following a verb as a direct object. Crucially, sentences like (2b) are coded as having a direct object (DO), from the learner's perspective. However, we excluded particles that make up a phrasal verb or simple adverbs from being considered as a direct object. For instance, for the verb 'pick', the utterance 'I picked up' or 'Did you pick up?' is coded as occurring without a direct object, even though the verb in question is followed by something other than punctuation.

In addition, each sentence is coded as being a question or not. We coded a sentence as a question if and only if the sentence occurs with a question mark in its transcript, which includes a lot of rising declaratives. Then, we defined pragmatics-augmented direct object (PDO) as 1 if and only if the sentence either has a DO or is a question, and 0 otherwise. The PDO coding is used as the input for the PUDL, which utilizes pragmatics, while the DO coding is used as the input for the distributional learner DL, not equipped with pragmatic knowledge.

The list of the 50 verbs with their total counts, sample DO rates, and sample PDO rates are shown in Table 1. Verbs are categorized according to their underlying true transitivity types following Perkins et al. (2022): (T)ransitive, (A)lternating, and (I)ntransitive. They are sorted by sample DO rates within each transitivity type. Transitive verbs tend to have higher sample DO rates and intransitive verbs tend to have lower sample DO rates. However, they can deviate much from the ground truth of 1 for transitive verbs and 0 for intransitive verbs. There is also a significant overlap of the sample DO rates among the three categories.

Finally, for each verb, its sample PDO rate is always higher than its sample DO rate as expected. For all the transitive verbs, the sample PDO rate is greater than 0.9, and one verb ('feed') attains a 100% sample PDO rate.

4 An empirical Bayes model for distributional learning

We propose an empirical Bayes (EB) model that conducts distributional learning of verb transitivity from observed DO patterns.

Model The model assumes that there are K transitivity categories $\{C_1, C_2, \ldots, C_K\}$ with equal prior weights. The transitivity T_i of each verb $i \in \{1, 2, \ldots, V\}$ is distributed as:

$$T_i \sim \text{Uniform}(\{C_1, C_2, \ldots, C_K\}).$$

Depending on transitivity category (C_k), the verb's true observable DO rate θ_i is drawn from an unknown Beta distribution (Beta(α_k, β_k)), taking values between 0 and 1:

$$\theta_i | T_i = C_k \sim \text{Beta}(\alpha_k, \beta_k).$$

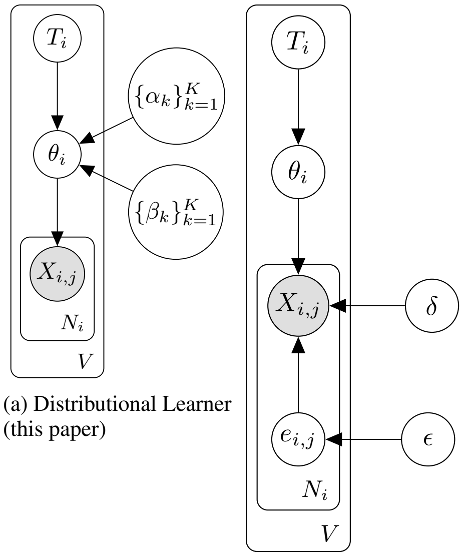
Verb	Count	DO Rate	PDO Rate
(T) feed	226	0.9690	1.0000
(T) hit	214	0.9579	0.9860
(T) bring	712	0.9424	0.9803
(T) throw	376	0.9282	0.9415
(T) fix	397	0.8992	0.9270
(T) buy	356	0.8989	0.9775
(T) hold	565	0.8690	0.9522
(T) catch	216	0.7731	0.9074
(T) wear	540	0.7241	0.9444
(A) pick	390	0.9410	0.9692
(A) drop	178	0.9157	0.9551
(A) knock	149	0.9128	0.9664
(A) touch	210	0.8857	0.9143
(A) push	348	0.8707	0.9282
(A) wash	236	0.8686	0.9576
(A) ride	243	0.8683	0.9630
(A) turn	470	0.8617	0.9277
(A) cut	318	0.8491	0.9403
(A) lose	200	0.8450	0.9000
(A) pull	383	0.8433	0.8799
(A) read	624	0.8301	0.8942
(A) leave	382	0.8246	0.8717
(A) build	307	0.8176	0.9479
(A) open	379	0.8153	0.8707
(A) bite	195	0.7949	0.9026
(A) close	212	0.7877	0.8491
(A) blow	214	0.7570	0.8738
(A) play	1424	0.7514	0.8820
(A) drink	345	0.7507	0.9420
(A) draw	401	0.7481	0.9202
(A) eat	1535	0.7036	0.8997
(A) sit	990	0.6939	0.8323
(A) move	260	0.6923	0.7846
(A) sing	347	0.6916	0.8646
(A) hang	168	0.6905	0.8690
(A) break	558	0.6900	0.7975
(A) write	593	0.6830	0.8499
(A) walk	255	0.6196	0.8078
(A) stand	300	0.5733	0.7800
(A) stick	278	0.5647	0.7626
(A) fit	211	0.5498	0.7536
(A) jump	189	0.5185	0.7354
(A) run	246	0.4837	0.7236
(A) swim	200	0.4500	0.7550
(I) wait	310	0.8452	0.8774
(I) stay	334	0.7575	0.8204
(I) sleep	419	0.4678	0.7709
(I) fall	606	0.3449	0.6188
(I) work	302	0.3377	0.5927
(I) cry	272	0.2647	0.6875

Table 1: Fifty verbs in our analysis with their total count, sample DO rate, and sample PDO rate.

Lastly, we assume that the DO observations $\{X_{i,j}\}_{j=1}^{N_i}$ are independently and identically distributed as a Bernoulli distribution with the success parameter equal to θ_i:

$$X_{i,j}|\theta_i \sim \text{Bernoulli}(\theta_i).$$

The left panel of Figure 1 summarizes our model in plate notation. Note that the verb's transitivity T_i and the verb's true observable DO rate θ_i are latent variables that need to be estimated, while the DO observation $X_{i,j}$ are observed variables (shaded in the Figure).



(a) Distributional Learner (this paper)

(b) Filtering Learner (Perkins et al., 2022)

Figure 1: Models in plate notation.

EB inference We have assumed that the model hyperparameters $\{(\alpha_k, \beta_k)\}_{k=1}^K$ are unknown. We estimate these hyperparameters using EB. Specifically, we set the hyperparameters to values that maximize the marginal log-likelihood.

The EB prior estimation and posterior computation can be done efficiently by reducing our model to the class of Beta-Binomial mixture models. We combine two simple observations: the marginal distribution of θ_i is a Beta mixture if we integrate T_i out; and the sum of the N_i Bernoulli trials is distributed as a Binomial distribution, $X_{i,\cdot} := \sum_{j=1}^{N_i} X_{i,j} \sim \text{Binomial}(N_i, \theta_i)$. Therefore, the sum of DO observations $X_{i,\cdot}$ is marginally distributed as a Beta-Binomial mixture:

$$X_{i,\cdot} \sim \frac{1}{K} \sum_{k=1}^{K} \text{Beta-Binomial}(N_i, \alpha_k, \beta_k).$$

We use the expectation–maximization (EM) algorithm to find the hyperparameter values that maximize this likelihood.

Initialization Since the likelihood maximization problem is not a convex problem, the solution obtained via the EM algorithm might depend on the initialization. We initialize the category memberships using a hard clustering of sample DO rates, $X_{i,\cdot}/N_i$.[2] For example, with $K = 3$ categories, we sort verbs by their sample DO rates, and assign a hard C_1/C_2/C_3 membership to the verbs with sample DO rates in the lowest/middle/upper tertile, respectively. The categories C_1, C_2, and C_3 are interpreted as the verb categories with 'low', 'middle', and 'high' true observable DO rates, which would roughly correspond to the 'intransitive', 'alternating', and 'transitive' categories of verb transitivity for the current problem.

Inference with PDO data To make inferences using the PDO data instead of the DO data, we use the same model and algorithm. The only difference is the interpretation of the model parameters: θ_i as the verb's true observable PDO rate and (α_k, β_k) as the parameters for the PDO distribution of the category C_k.

5 Results

We use the EB model to simulate a distributional learner (DL) that learns verb transitivity from DO data, and a pragmatics-utilizing distributional learner (PUDL) that learns verb transitivity from pragmatics-augmented DO (PDO) data, which incorporates pragmatic knowledge about questions.

5.1 Distributional Learner (DL)

To simulate a DL, we fit the EB model with three categories ($K = 3$), consistent with the underlying truth that there are three verb transitivity categories (intransitive, alternating, and transitive).

The estimated hyperparameters (α_k, β_k) for the EB Beta priors are $(4.76, 3.64)$, $(28.58, 8.60)$, and $(33.39, 4.28)$ for the categories C_1, C_2, and C_3; their means are 0.57, 0.77, and 0.89. The densities of the three distributions are shown in the uppermost panel of Figure 2. Note that we do not use the

[2] In a hard clustering, each verb i belongs to only one category, whereas, in a soft clustering, it can belong to multiple categories. It is worth noting that the hard clustering-based initialization is an initialization strategy, not a part of the model specification, though the initialization can have a lasting impact on the final inference.

true verb transitivity labels (‘intransitive’, ‘alternating’, or ’transitive’) in the estimation procedure.

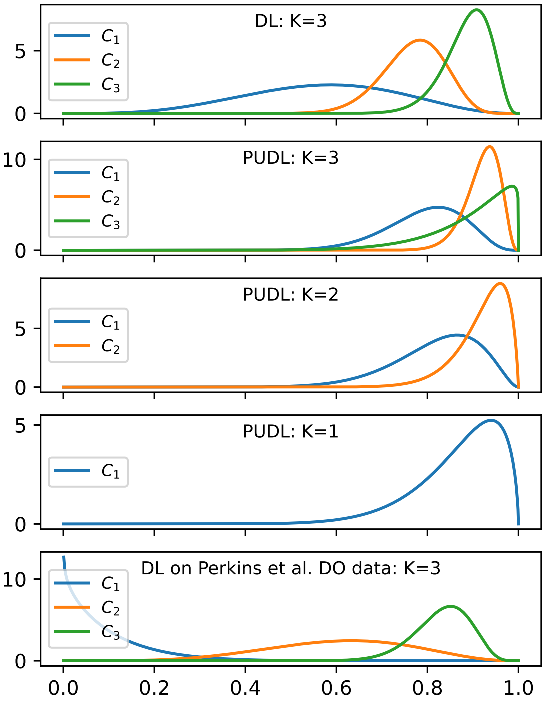


Figure 2: Empirical Bayes beta priors.

Based on the empirical Bayes beta priors, each verb’s posterior distribution over the verb categories $C_1, \dots, C_K$ is computed. Each verb’s posterior membership in the categories is shown as a stacked bar plot in the uppermost panel of Figure 3; the posterior memberships are non-negative and sum to one. The verb labels in the x-axis are color-coded to represent the underlying true transitivity category: transitive verbs are coded red, alternating verbs are coded black, and intransitive verbs orange. The verbs are ordered first by the underlying true transitivity category, and then by the descending sample DO rate within each category.

Our EB model performs well in uncovering the underlying true transitivity category, though not perfectly. Out of the nine transitive verbs, seven verbs have the highest membership in the ‘high’ category C_3, which is the category with the highest prior DO rates; the other two transitive verbs have the highest membership in the ‘middle’ category C_2. On the other side, four out of the six intransitive verbs have the highest membership in the ‘low’ category C_1. The alternating verbs have varying levels of memberships in the three categories, depending on their sample DO rates.

5.2 Pragmatics-Utilizing Distributional Learner (PUDL)

To simulate a PUDL, we fit the EB model with three categories ($K = 3$) to the PDO data. The estimated EB beta priors and the posterior memberships are shown in the second uppermost panels of Figure 2 and 3. In Figure 3, verbs within each category are reordered according to their sample PDO rates. Compared to the DL, the PUDL has verbs’ posterior memberships less separated. For example, all the fifty verbs have non-negligible memberships in the C_3 category, and the transitive verbs’ C_3 membership decreased. This change follows from the property of the PDO data: each verb’s PDO rate is always greater than or equal to its DO rate, and the verbs’ PDO rates are harder to separate into distinct clusters, since they are all shifted toward 1 (closer to 1 than the DO rates are). This property is illustrated in the estimated EB beta priors in the second panel of Figure 2, which is more overlapping than the first panel.

We find that the PUDL favors models with a smaller number of categories, based on the Bayesian Information Criterion (BIC). BIC is a criterion for model selection, which is defined as

$$\mathrm{BIC} = -2\log(\hat{L}) + P\log(N)$$

where $\hat{L}$ is the maximized log-likelihood of the model, P is the number of parameters estimated by the model, and N is the sample size. A model with a smaller BIC is preferred. To strike a balance between model fit and model complexity, BIC adds a penalty to the number of parameters, as models with a larger number of parameters are more flexible to guarantee a higher maximized log-likelihood.

K	BIC	$-2\log(\hat{L})$	$P\log(N)$
✓1	478.5759	470.7519	7.8240
2	486.3513	470.7032	15.6481
3	493.5469	470.0747	23.4721

Table 2: Bayesian Information Criterion for PUDL.

In our case, the sample size N is 50 and the number of parameters P is $2K$ from the size of the set $\{(\alpha_k, \beta_k)\}_{k=1}^{K}$. The PUDL with $K = 3$ has BIC 493.55, BIC 486.35 with $K = 2$, and 478.58 with $K = 1$ (see Table 2). Therefore, the PUDL with $K = 1$ is the most preferred, and the PUDL with $K = 3$ is the least preferred, among the three models. The estimated EB prior for the PUDL with

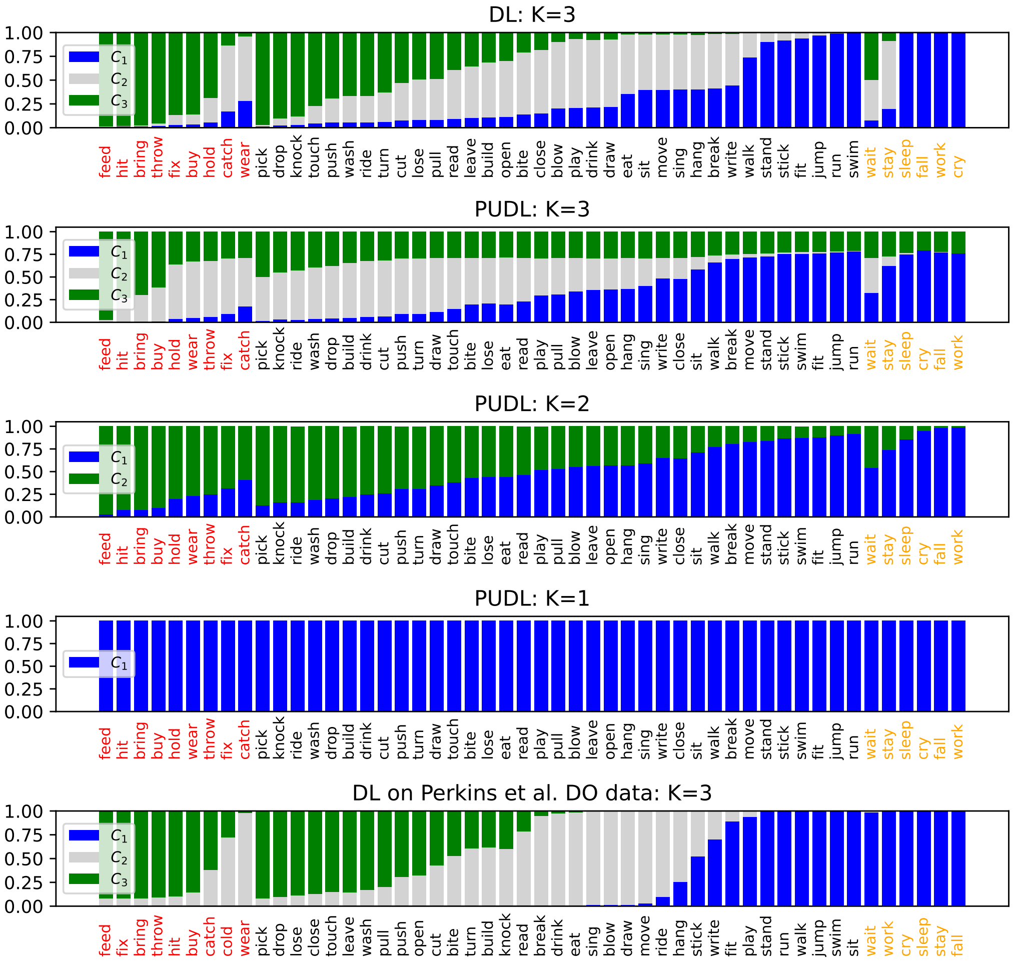


Figure 3: Posterior distributions over verb categories T.

$K = 1$ is shown in the second lowermost panel of Figure 2; its true observable PDO rates are concentrated around large values. Naturally, the posterior membership for each verb is 1 in the only available category C_1, as shown in the second lowermost panel of Figure 3. For completeness, the estimated EB prior and posterior memberships for $K = 2$ are provided in the middle panel of Figure 2 and 3, respectively.

Intuitively, the preference for $K = 1$ indicated by BIC suggests that the pragmatically-informed learner infers that the observations are coming from a single common source, rather than two or three clusters. Capable of pragmatic reasoning about the question-answer relation, the learner made an impressive progress by recognizing verbs like 'throw' in (1) are more transitive than it previously thought they would be. However, the learner at the assumed developmental stage is still potentially misguided by the data like (2b) for intransitive verbs, making an incorrect inference that verbs like 'wait' can occur with a direct object. Consequently, the learner undergoes the overregularizing stage, where it perceives all verbs as belonging to one category, i.e., a category with high true observable DO rates.[3] This explains why $K = 1$ is preferred when the model is asked to cluster 50 verbs into K-many categories. Once the learner resolves the PP vs. NP distinction at a later stage of development, possibly after 19 months old given the experimental results in Lidz et al. (2017), we expect the result for the PUDL $K = 3$ to be more clearly separated than the DL $K = 3$, showing more progress toward deterministic knowledge. We leave the experimentation with the PUDL augmented by the PP vs. NP resolution for future research.

[3] It is possible to interpret this single category as either transitive or alternating. The upshot is that the learner would infer that verbs are followed by a direct object with a high probability.

6 Comparison with a filtering model

In this section, we compare our distributional learner with a filtering-based distributional learner, proposed by Perkins et al. (2022).

Filtering model The filtering-based learner identifies and filters out the inherent noise in the overt DO data, such as (1b). Assuming deterministic hypotheses of 0% DO rate, 100% DO rate, and 0-100% DO rate for intransitive, transitive, and alternating categories, respectively, the model incorporates filtering as inductive bias, allowing it to arrive at accurate generalizations only by looking at the rates of overt objects following verbs. What sets this approach apart from other proposals on filtering is that the learner operates without predetermined understanding of which data is misleading in terms of verb transitivity. All it assumes is a certain amount of noise in the data, acknowledging the presence of erroneous parses. The key insight of Perkins et al. (2022) is that the learner confronts the complex transitivity learning problem by filtering out these erroneous parses without necessarily knowing that the data such as (1b) and (2b) are non-basic clauses.

The filtering learner assumes that there are three transitivity categories $\{C_t, C_a, C_i\}$ (transitive, alternating, and intransitive) with equal prior weights. The transitivity T_i of each verb $i \in \{1, 2, \ldots, V\}$ is distributed as:

$$T_i \sim \text{Uniform}(\{C_t, C_a, C_i\}).$$

Depending on the transitivity category, the verb's true DO rate θ_i is drawn from known deterministic values or a known distribution:

$$\theta_i | T_i \sim \begin{cases} \delta_{(1)}, & \text{if } T_i = C_t \\ \text{Uniform}([0,1]), & \text{if } T_i = C_a \\ \delta_{(0)}, & \text{if } T_i = C_i \end{cases}$$

This modeling choice encodes the deterministic hypothesis space in which there is a known category that always has a DO ('transitive') and another known category that never has a DO ('intransitive'). By contrast, the categories in our model have DO rates from a flexible Beta distribution, not tied to specific values.

Notice also the difference in the definition of the parameter θ_i as a verb's *true DO rate* in their modeling versus a verb's *true observable DO rate* in ours. The reason behind our learner's modeling true observable DO rate, not true DO rate, is because our learner does not have prior knowledge about the deterministic hypotheses. Our learner is purely distributional; all the input they receive, including the utterances that we described as "misleading" above, are potentially signals that drive transitivity learning. In this regard, it is *true observable DO rate*, not *true DO rate*.

On the other hand, the filtering learner explicitly models the "misleading" part of data as noise. First, there is a parameter ϵ for the probability of an erroneous parse, which is distributed as a uniform distribution:

$$\epsilon \sim \text{Uniform}([0,1]).$$

Second, there is another parameter δ for the probability of generating a DO in error, which is distributed as a uniform distribution:

$$\delta \sim \text{Uniform([0,1])}.$$

Third, there is a sentence-level "input filter" $e_{i,j}$; $e_{i,j} = 1$ means the observation $X_{i,j}$ is generated from erroneous parsing. The input filter is modeled as a Bernoulli distribution:

$$e_{i,j} | \epsilon \sim \text{Bernoulli}(\epsilon).$$

Lastly, the overt DO observation $X_{i,j}$ is modeled as a mixture of the two Bernoulli distributions with success probability θ_i and δ. $X_{i,j} = 1$ means the sentence j of verb i has a DO.

$$X_{i,j} | \delta, \theta_i, e_{i,j} \sim \begin{cases} \text{Bernoulli}(\theta_i), & \text{if } e_{i,j} = 0 \\ \text{Bernoulli}(\delta), & \text{if } e_{i,j} = 1. \end{cases}$$

The filtering-based model is illustrated in the right panel of Figure 1.

Data For comparison, we present our DL's performance on the DO data reported in Perkins et al. (2022). Note that although we follow their list of fifty verbs and use the same corpora in our analysis, the exact total count and sample DO rates are different. Specifically, the DO rates tend to be higher in our dataset because we assume that our learner hasn't yet resolved the NP vs. PP distinction. By contrast, Perkins et al. (2022) define the overt DO as "right NP sisters of V", which suggests that their learner can distinguish PPs from NP objects.

Result Our DL's estimated hyperparameters (α_k, β_k) for EB Beta priors are $(0.91, 8.91)$, $(5.66, 3.70)$, and $(30.39, 6.14)$ for the categories C_1, C_2, and C_3. The means are 0.09, 0.60, and 0.83; the densities are shown in the lowermost panel of Figure 2, and each verb's posterior memberships in the lowermost panel of Figure 3.

We find that our posterior membership results closely align with Figure 2 of Perkins et al. (2022). The successful verb transitivity learning reported in Perkins et al. (2022) has been attributed to the filtering mechanism, a type of inductive bias that enforces a deterministic hypothesis space. Our learner, in contrast, does not entertain a restricted hypothesis space to start with, which suggests that pure distributional learning is enough to replicate successful transitivity learning. We also highlight that our learning algorithm is simpler and more efficient than the filtering algorithm, with the runtime being less than a second.

7 Conclusion

We introduced Pragmatics-Utilizing Distributional Learner (PUDL) to model verb transitivity learning, assuming the grammatical knowledge typical of 15-month-old English learners. PUDL integrates learners' pragmatic reasoning, particularly the realization that utterances such as 'What did Alex throw?' are information-seeking questions, leading in turn to the inference that this type of object wh-questions would lack a direct object, i.e., the *answer* to the question being asked. These neutral object wh-questions do not confuse pragmatically informed learners of verb transitivity, even though 'throw', in principle, is a transitive verb that requires a direct object. The nuanced pragmatic reasoning prompts learners to adjust their initial generalization closer to adult grammar in the domain of transitive verbs. However, the proposed pragmatic knowledge alone is insufficient to handle the noisy data in the domain of intransitive verbs. Specifically, we predicted a developmental trajectory characterized by a temporary overregularization stage, where learners generalize all verbs into a single category in terms of transitivity due to difficulty in distinguishing PP adjuncts from NP arguments. Once the PP and NP distinction is resolved[4], possibly after 19 months of age, as suggested by Lidz et al.'s (2017) behavioral studies, we anticipate the resolution of overgeneralization and significant progress in the intransitive domain as well, which we leave for future research. It remains to be demonstrated by behavioral experiments whether children at this critical period indeed exhibit overregularization, categorizing both the transitive verb 'throw' and the intransitive verb 'wait' into the same category in terms of transitivity. Nevertheless, we have shown that the proposed purely distributional models, Distributional Learner (DL) and Pragmatics-Utilizing Distributional Learner (PUDL), which operate confidently with received data, are as promising as an alternative distributional model that considers deterministic hypothesis space and filters out a portion of input as noise.

Acknowledgements

We thank Laurel Perkins for her invaluable guidance and discussions, which were instrumental in shaping this work, and the three anonymous reviewers for their constructive and helpful feedback. We are also grateful to the participants of the Fall 2023 UCLA graduate course on grammatical development for inspiring discussions on learning.

References

Alison C Austin, Kathryn D Schuler, Sarah Furlong, and Elissa L Newport. 2022. Learning a language from inconsistent input: Regularization in child and adult learners. *Language Learning and Development*, 18(3):249–277.

Leon Bergen, Edward Gibson, and Timothy J O'Donnell. 2022. Simplicity and learning to distinguish arguments from modifiers. *Journal of Language Modelling*, 10.

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press.

Daniel Buring and Christine Gunlogson. 2000. Aren't positive and negative polar questions the same?

Donka F Farkas and Floris Roelofsen. 2017. Division of labor in the interpretation of declaratives and interrogatives. *Journal of Semantics*, 34(2):237–289.

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):234–260.

[4] For instance, see Bergen et al. (2022) for recent computational modeling work on how learners differentiate between arguments and adjuncts based on distributional information. The current paper does not depend on exactly which model is adopted for the NP argument-PP adjunct resolution, as behavioral studies suggest that this differentiation is observed at least after 19 months of age, which is later than the developmental stage assumed throughout this paper.

Carla L Hudson Kam and Elissa L Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, 1(2):151–195.

Jeffrey Lidz, Aaron Steven White, and Rebecca Baier. 2017. The role of incremental parsing in syntactically conditioned word learning. *Cognitive Psychology*, 97:62–78.

Brian MacWhinney. 2000. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.

Laurel Perkins, Naomi H Feldman, and Jeffrey Lidz. 2022. The power of ignoring: filtering input for argument structure acquisition. *Cognitive Science*, 46(1):e13080.

Laurel Perkins and Jeffrey Lidz. 2021. Eighteen-month-old infants represent nonlocal syntactic dependencies. *Proceedings of the National Academy of Sciences*, 118(41):e2026469118.

Melanie Soderstrom, Megan Blossom, Rina Foygel, and James L Morgan. 2008. Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, 35(4):869–902.

Patrick Suppes. 1974. The semantics of children's language. *American psychologist*, 29(2):103.

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