Learning Verb Classes in an Incremental Model

Libby Barak, Afsaneh Fazly, and **Suzanne Stevenson**
Department of Computer Science
University of Toronto
Toronto, Canada
{libbyb,afsaneh,suzanne}@cs.toronto.edu

Abstract

The ability of children to generalize over the linguistic input they receive is key to acquiring productive knowledge of verbs. Such generalizations help children extend their learned knowledge of constructions to a novel verb, and use it appropriately in syntactic patterns previously unobserved for that verb—a key factor in language productivity. Computational models can help shed light on the gradual development of more abstract knowledge during verb acquisition. We present an incremental Bayesian model that simultaneously and incrementally learns argument structure constructions and verb classes given naturalistic language input. We show how the distributional properties in the input language influence the formation of generalizations over the constructions and classes.

1 Introduction

Usage-based accounts of language learning note that young children rely on verb-specific knowledge to produce their early utterances (e.g., Tomasello, 2003). However, evidence suggests that even young children can generalize their verb knowledge to novel verbs and syntactic frames (e.g., Fisher, 2002), and that the abstract knowledge gradually strengthens over time (e.g., Tomasello and Abbot-Smith, 2002). One area of verb usage where more sophisticated abstraction appears necessary for fully adult productivity in language is the knowledge of verb alternations. A verb alternation is a pairing of constructions shared by a number of verbs, in which the two constructions express related argument structures (Levin, 1993): e.g., the dative alternation involves the related forms of the prepositional dative (PD; *X gave Y to Z*) and the double-object dative (DO; *X gave Z Y*). Such alternations enable language users to readily adapt new and low frequency verbs to appropriate constructions of the language by generalizing the observed use of one such form to the other.[1]

For example, Conwell and Demuth (2007) show that 3-year-old children understand that a novel verb observed only in the DO dative (*John gorped Heather the book*) can also be used in the PD form (*John gorped the book to Heather*), though the children can only generalize such knowledge under certain experimental conditions. Wonnacott et al. (2008) demonstrate the proficiency of adults in making such generalizations within an artificial language learning scenario, which enables the researchers to explore the distributional properties of the linguistic input that facilitate the acquisition of such generalizations. The results suggest that the overall frequency of the syntactic patterns as well as the distribution of verbs across the patterns play a facilitatory role in the formation of abstract verb knowledge (in the form of verb alternations) in adult language learners.

In this work, we propose a computational model that extends an existing Bayesian model of verb argument structure acquisition (Alishahi and Stevenson, 2008)[AS08] to support the learning of verb classes over the acquired constructions. Our model is novel in its approach to verb class formation, because it clusters tokens of a verb that reflect the distribution of the verb over the learned constructions each time the verb is used in an input. That is, the model forms verb classes by clustering verb tokens that reflect the evolving usages of the verbs in various constructions.

We use this new model to analyze the role of the classes and the distributional properties of the input in learning abstract verb knowledge, given

[1] The *generalization of an alternation* refers to a speaker using one variant of an alternation for a verb (e.g., PD) having only observed the verb in the other variant (e.g., DO).

naturalistic input that contains many verbs and many constructions. The model can form higher-level generalizations such as learning verb alternations, which is not possible with the AS08 model (cf. the findings of Parisien and Stevenson, 2010). Moreover, because our model gradually forms its representations of constructions and classes over time (in contrast to other Bayesian models, such as Parisien and Stevenson, 2010; Perfors et al., 2010), it is possible to analyze the monotonically-growing representations and show their compatibility with the developmental patterns seen in children (Conwell and Demuth, 2007). We also replicate some of the observations of Wonnacott et al. (2008) on the role of distributional properties of the language in influencing the degree of generalization over an alternation.

2 Related Work

To explore the properties of learning mechanisms that are capable of mimicking child and adult psycholinguistic observations, a number of cognitive modeling studies have focused on learning abstract verb knowledge from individual verb usages (e.g., Alishahi and Stevenson, 2008; Perfors et al., 2010; Parisien and Stevenson, 2010). Here we focus on such computational models that enable the sort of higher-level generalization that people do across verb alternations, unlike the AS08 model.

The hierarchical Bayesian models of Perfors et al. (2010) and Parisien and Stevenson (2010) focus on learning this kind of higher-level generalization. The model of Perfors et al. (2010) learns verb alternations, i.e., pairs of syntactic patterns shared by certain groups of verbs. By incorporating this sort of abstract knowledge into their model, Perfors et al. are able to simulate the ability of adults to generalize across verb alternations (as in Wonnacott et al., 2008). That is, Perfors et al. predict the ability of a novel verb to occur in a syntactic structure after exposure to it in the alternative pattern of that alternation. However, this model is trained on data that contains only a limited number of verbs and syntactic patterns unlike naturalistic Child-directed Speech (CDS) and moreover incorporates built-in information about verb constructions.

The hierarchical Dirichlet model of Parisien and Stevenson (2010) addresses these limitations by working with natural child-directed speech (CDS) data. Moreover, the model of Parisien and Stevenson simultaneously learns constructions as in AS08 and verb classes based on verb alternation behaviour, showing that the latter level of abstraction is necessary to support effective learning of verb alternations. Still, the models of both Parisien and Stevenson and Perfors et al. can only be utilized as a batch process and hence are limited in the analysis of developmental trajectories. Although it is possible to simulate development by training such models on increasing portions of input, such an approach does not ensure that the representations given $n + i$ inputs can be developed from the representation given n inputs.

In this paper, we propose a significant extension to the model of AS08, by adding an extra level of abstraction that incrementally learns verb classes by drawing on the distribution of verbs over the learned constructions. The new model combines the advantages of having a monotonic clustering model that enables the analysis of developing clusters, with the simultaneous learning of constructions and verb classes.

3 The Computational Model

As mentioned above, our model is an extension of the model of AS08 in which we add a level of learned abstract knowledge about verbs. Specifically, our model uses a Bayesian clustering process to learn clusters of verb usages that occur in similar argument structure constructions, as in the original model of AS08. To this, we add another level of abstraction that learns clusters of verbs that exhibit similar distributional patterns of occurrence across the learned constructions—that is, classes of verbs that occur in similar *sets of* constructions, and in similar proportions. To distinguish between the clusters of the two levels of abstraction in our new model, we refer to the clusters of verb usages as constructions, and to the groupings of verbs given their distribution over those constructions as verb classes.

3.1 Overview of the Model

The model learns from a sequence of *frames*, where each frame is a collection of *features* representing what the learner might extract from an utterance s/he has heard. Similarly to previous computational studies (e.g., Parisien and Stevenson, 2010), here we focus on syntactic features since our goal is to understand the acquisition of acceptable syntactic structures of verbs indepen-

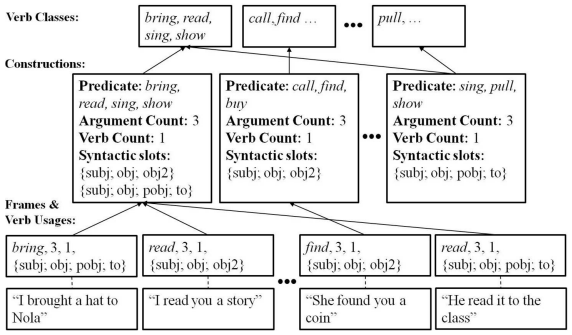


Figure 1: A visual representation of the two levels of abstraction in the model, with sample verb usages input (and extracted input frames), constructions, and classes.

dently of their meaning, as in some relevant psycholinguistic (Wonnacott et al., 2008) and computational studies (Parisien and Stevenson, 2010). We focus particularly on properties such as *syntactic slots* and *argument count*. (These features, as in Parisien and Stevenson (2010), provide a more flexible and generalizable representation of a syntactic structure than the syntactic pattern string used by AS08.) See the bottom rows of boxes in Figure 1 for sample input verb usages with their extracted frames.

The model incrementally clusters the extracted input frames into *constructions* that reflect probabilistic associations of the features across similar verb usages; see the middle level of Figure 1. Each learned cluster is a probabilistic (and possibly noisy) representation of an argument structure construction: e.g., a cluster containing frames corresponding to usages such as *I eat apples*, *She took the ball*, and *He got a book*, etc., represents a Transitive Action construction.[2] Such constructions allow for some degree of generalization over the observed input; e.g., when seeing a novel verb in a Transitive utterance, the model predicts the similarity of this verb to other Action verbs appearing in that pattern (Alishahi and Stevenson, 2008).

Grouping of verb usages into constructions may not be sufficient for making higher-level generalizations across verb alternations. Knowledge of alternations is only captured indirectly in constructions (because usages of the same verb can occur in multiple clusters). Following Parisien and Stevenson (2010), we hypothesize that true generalization behaviour requires explicit knowledge that verbs have commonalities in their patterns of occurrence *across* constructions; this is the basis for verb classes (Levin, 1993; Merlo and Stevenson, 2000; Schulte im Walde and Brew, 2002).

To capture this, our model learns groupings of verbs that have similar distributions across the learned constructions. These groupings form verb classes that provide a higher-level of abstraction over the input; see the top level in Figure 1. Consider the dative alternation: the classes capture the fact that some verbs may occur only in prepositional dative (PD) forms, such as *sing*, while others occur only in double object (DO) forms (*call*), while still others *alternate* – i.e., they occur in both (*bring*).

Our model simultaneously learns both of these types of knowledge: constructions are clusters of verb usages, and classes are clusters of verb distributions over those constructions. Importantly, it does so incrementally, which allows us to examine the developmental trajectory of acquiring alternations such as the dative as the learned clusters grow over time. Moreover, both types of clustering are monotonic, i.e., we do not re-structure the groupings that our model learns. However, the model in both levels is clustering *verb tokens* – i.e., the features corresponding to the verb at that time in the input, its usage or its current distribution – so that the same verb type may be added to various clusters at different stages in the training.

3.2 Learning Constructions of Verb Usages

The model of AS08 groups input frames into clusters on the basis of the overall similarity in the values of their features. Importantly, the model learns these clusters incrementally in response to the input; the number and type of clusters is not predetermined. The model considers the creation of a new cluster for a given frame if the frame is not sufficiently similar to any of the existing clusters. Formally, the model finds the best cluster for a given input frame F as in:

$$\text{BestCluster}(F) = \underset{k \in Clusters}{\text{argmax}}\ P(k|F) \quad (1)$$

where k ranges over all existing clusters and a new one. Using Bayes rule:

$$P(k|F) = \frac{P(k)P(F|k)}{P(F)} \propto P(k)P(F|k) \quad (2)$$

The prior probability of a cluster $P(k)$ is estimated as the proportion of frames that are in k out of all observed input frames, thus assigning a higher

[2] Because the associations are probabilistic, a linguistic construction may be represented by more than one cluster.

prior more frequent constructions. The likelihood $P(F|k)$ is estimated based on the match of feature values in F and in the frames of k (assuming independence of the features):

$$P(F|k) = \prod_{i \in Features} P_i(j|k) \quad (3)$$

where j is the value of the i^{th} feature of F, and $P_i(j|k)$ is calculated using a smoothed version of:

$$P_i(j|k) = \frac{\text{count}_i(j,k)}{n_k} \quad (4)$$

where $\text{count}_i(j,k)$ is the number of times feature i has the value j in cluster k, and n_k is the number of frames in k. We compare the slot features as sets to capture similarities in overlapping syntactic slots rather than enforcing an exact match. The model uses the Jaccard similarity score to measure the degree of overlap between two feature sets, instead of the direct count of occurrence in Eqn. (4):

$$\text{sim\_score}(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \quad (5)$$

where S_1 and S_2 in our experiments here are the sets of syntactic slot features.

3.3 Learning Verb Classes

Our new model extends the construction-formation model of AS08 by grouping verbs into classes on the basis of their distribution across the learned constructions. That is, verbs that have statistically-similar patterns of occurrence across the learned constructions will be considered as forming a verb class. For example, in Figure 1 we see that *bring* and *read* may be put into the same class because they both occur in a similar relative frequency across the DO and PD constructions (the leftmost and rightmost constructions in the figure).

We use the same incremental Bayesian clustering algorithm for learning the verb classes as for learning constructions. At the class level, the feature used for determining similarity of items in clustering is the distribution of each verb across the learned constructions. As for constructions, the model learns the verb classes incrementally; the number and type is not predetermined. Moreover, just as constructions are gradually formed from successively processing a particular verb usage at each input step, the model forms verb classes from a sequence of snapshots of the input verb's distribution over the constructions at each input step. This means that our model is forming classes of verb tokens rather than types; if a verb's behaviour changes over the duration of the input, subsequent tokens (the distributions over constructions at later points in time) may be clustered into a different class (or classes) than earlier tokens, even though prior decisions cannot be undone.

Formally, after clustering the input frame at time t into a construction, as explained above, the model extracts the current distribution d_{v_t} of its head verb v over the learned constructions; this is estimated as a smoothed version of v's relative frequency in each construction:

$$P(k|v) = \frac{\text{count}(v,k)}{n_v} \quad (6)$$

where $\text{count}(v,k)$ is the number of times that inputs with verb v have been clustered into construction k, and n_v is the number of times v has occurred in the input thus far.

To cluster this snapshot of the verb's distribution, d_{v_t}, it is compared to the distributions encoded by the model's classes. The distribution d_c of an existing class c is the weighted average of the distributions of its member verb tokens:

$$d_c = \frac{1}{|c|} \sum_{v \in c} \text{count}(v,c) \times d_v \quad (7)$$

where $|c|$ is the size of class c, $\text{count}(v,c)$ is the number of occurrences of v that have been assigned to c, and d_v is the distribution of the verb v given by the tokens of v (the "snapshots" of distributions of v assigned to class c). That is, d_v in c is an average of the distributions of all d_{v_t} for verb v that have been clustered into c.

The model finds the best class for a given verb distribution d_{v_t} based on its similarity to the distributions of all existing classes and a new one:

$$\text{BestClass}(d_{v_t}) = \underset{c \in Classes}{\text{argmax}} \; (1 - D_{\text{JS}}(d_c \| d_{v_t})) \quad (8)$$

where c ranges over all existing classes as well as a new class that is represented as a uniform distribution over the existing constructions. Jensen–Shannon divergence, D_{JS}, is a popular method for measuring the distance between two distributions: It is based on the KL–divergence, but it is symmetric and has a finite value between 0 and 1:

$$D_{\text{JS}}(p\|q) = \frac{1}{2} D_{\text{KL}}(p\|\frac{1}{2}(p+q)) + \frac{1}{2} D_{\text{KL}}(q\|\frac{1}{2}(p+q)) \quad (9)$$

	non-ALT		ALT	
	DO-only	PD-only	DO	PD
Number of verbs	12	5	6	
Relative frequency	14%	2%	2%	1%

Table 1: Number of non-alternating (non-ALT) and alternating (ALT) verbs in our lexicon, as well as the relative frequency of each construction in our generated input corpora.

4 Experimental Setup

4.1 Generation of the Input Corpora

We follow the input generation method of AS08 to create naturalistic corpora that are based on the distributional properties of verbs over various constructions, as observed in child-directed speech (CDS). Our *input-generation lexicon* contains 71 verbs drawn from AS08 (11 action verbs) and Barak et al. (2013) (31 verbs of varying syntactic patterns), plus an additional 40 of the most frequent verbs in CDS, in order to have a range of verbs that occur with the PD and DO constructions. Table 4.1 shows the number of verbs that appear in the DO or PD construction only (non-alternating), as well as those that alternate across the two. (The table also gives the relative frequency of each dative construction in our generated input corpora.) Each verb lexical entry includes its overall frequency, and its relative frequency with each of a number of observed syntactic constructions. The frequencies are extracted from a manual annotation of a sample of 100 child-directed utterances per verb from a collection of eight corpora from CHILDES (MacWhinney, 2000).[3] An input corpus is generated by iteratively selecting a random verb and a syntactic construction based on their frequencies according to the lexicon, so that all input corpora used in our simulations have the distributional properties observed in CDS, but show some variation in precise make-up and ordering of verb usages. The generated input consists of *frames* (a set of features) that correspond to verb usages in CDS.

4.2 Simulations

Because the generation of the input data is probabilistic, we conduct 100 simulations for each experiment (each using a different input corpus) to avoid any dependency on specific idiosyncratic properties of a single generated corpus. For each simulation, we train our model on an automatically-generated corpus of $15,000$ frames, from which the model learns constructions and verb classes. At specified points in the input, we present the model with usages of a novel verb in a DO and/or PD frame, and then test the model's generalization ability by predicting DO and PD frames given that verb. Since we are interested in the relative likelihoods of the two frames, we report the difference between the log-likelihood of the DO frame and the log-likelihood of the PD frame, i.e., $\text{log-likelihood}(\text{DO}) - \text{log-likelihood}(\text{PD})$.

Specifically, we form a partial frame F_{test} (containing all usage features except for the verb) that reflects either the PD or the DO syntax, and assess the probability $P(F_{\text{test}}|v)$ for each of these, as in:

$$P(F_{\text{test}}|v) = \sum_{k \in Constructions} P(F_{\text{test}}|k)P(k|v) \tag{10}$$

where $P(F_{\text{test}}|k)$ is calculated as in Eqn. (3).

We can calculate $P(k|v)$ in two different ways: using only the knowledge in the constructions of the model, and using the knowledge that takes into account the verb classes over the constructions. For model predictions based on the construction level only, we calculate $P(k|v)$ as in Eqn. (6), which is the smoothed relative frequency of the verb v over construction k.

Predictions using knowledge of the verb classes will instead determine $P(k|v)$ drawing on the fit of verb v to the various classes (specifically, the similarity of v's distribution over constructions to the distribution encoded in each class), and the likelihood of each construction k for each class c (specifically, the likelihood of k given the distribution over constructions encoded in c), as in:

$$P(k|v) \approx \sum_{c \in Classes} P(k|c)P(c|v) \tag{11}$$

where $P(k|c)$ is the probability of construction k given class c's distribution over constructions (d_c); and $P(c|v)$ is the probability of c given verb v's distribution d_v over the constructions (using Jensen-Shannon divergence as in Eqn. (9)).

Due to the different number of clusters in each of the construction and class layers of the model, the likelihoods computed for each will differ in the range of values. For this reason, specific values cannot be directly compared across the layers of the model, rather we must analyze the general trends of the construction-only and class-based results.

[3] Brown (1973); Suppes (1974); Kuczaj (1977); Bloom et al. (1974); Sachs (1983); Lieven et al. (2009).

5 Evaluation

In this section we examine whether and how our model generalizes across the two variants of the dative alternation, the double-object dative (DO) and the prepositional dative (PD). To do so, we measure the tendency of the model to produce a novel verb observed in one dative frame in that same frame, or in the other dative frame (unobserved for that verb). Our goal is to understand the impact of the learned constructions and classes on this generalization behaviour. Following Parisien and Stevenson (2010), we examine three input conditions in which the novel verb occurs: (i) twice with the DO syntax (non-alternating); (ii) twice with the PD syntax (non-alternating); or (iii) once each with DO and PD syntax (alternating).[4] We then ask the model to predict the likelihood of producing each dative frame with that verb. Our focus here is on comparing the generalization abilities of the two levels of abstract knowledge in our model: the constructions versus the verb classes.

As a reminder, we use the dative alternation as one example for considering this kind of higher-level generalization behaviour observed in adults and to a lesser extent in children. Moreover, we perform the analysis in the context of naturalistic input that contains many verbs (those that appear in the dative and those that do not), and a variety of constructions , to provide a realistic setting for the task. Our settings differ from the psycholinguistic studies in the variability of constructions compared with the artificial language used by Wonnacott et al., and in focusing only on the syntactic properties unlike Conwell and Demuth. However, we follow the settings of these studies in analyzing the syntactic properties of a generated utterance given minimal exposure to a novel verb. Therefore, we aim to replicate their general observations by showing that (i) children are limited in their ability to generalize across verb alternations compared with adults, and (ii) the frequency of a construction has a positive correlation with the generalization rate of the construction.

5.1 Generalization of Learned Knowledge

We examine the generalization patterns of our model when presented with a novel verb in DO/PD forms after being trained on $15,000$ inputs, which we compare to the performance of adults in such language tasks. We first consider the case where the model predictions are based solely on the knowledge of constructions. Here we expect the predictions to correspond to the syntactic properties of the two inputs observed for the novel verb, with limited generalization. That is, we expect a non-alternating verb to be much more likely in the observed dative frame, and an alternating verb to be equally likely in both frames. The left hand side of Figure 2 presents the differences in log-likelihoods of the predicted DO and PD frames for the novel verb using the construction-based probabilities. The results confirm our expectation that the knowledge of constructions can support only limited generalization *across* the variants of an alternation. For the non-alternating conditions, the observed frame is highly favoured, and for the Alternating test scenario, the DO and PD frames have nearly equal likelihoods.

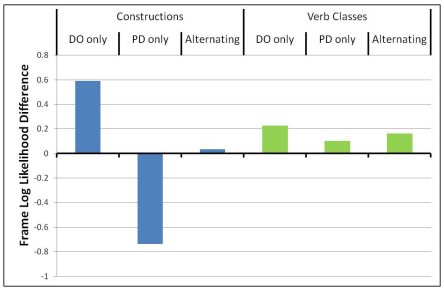


Figure 2: The difference between the log-likelihood values of the DO and PD frames, given each of the three input conditions: DO only, PD only, and Alternating. Values above zero denote a higher likelihood for the DO frame, and values below zero denote a higher likelihood for the PD frame.

We next turn to using the knowledge of verb classes, which we expect to enable generalizations that correspond to verb alternation behaviour — that is, we expect the model predictions here to reflect the knowledge that verbs that occur in one form of the alternation also often occur in the other form of the alternation. This is possible because the classes in the model encode the distributional patterns of verbs across constructions. In the absence of other factors, we would expect the Alternating condition to again show near equal likelihoods for the two frames, and the two non-alternating conditions to show a slight preference for the observed frame (rather than the strong preference seen in the construction-based predictions), because the unobserved frame is also likely due to the knowledge here of the alternation.

The right hand side of Figure 2 presents the

[4] For the alternating condition, half the simulations have DO first, and half have PD first.

difference in the log-likelihoods of the DO and PD frames when using the knowledge encoded in the verb classes. The results are not directly in line with the simple prediction above: The non-alternating (DO-only and PD-only) conditions show a weak preference (as expected) for one frame over another, but both favour the DO frame, as does the Alternating condition. That is, the PD-only and Alternating conditions show a preference for the DO frame that does not follow simply from the knowledge of alternations.

The DO preference in the PD-only and Alternating conditions arises due to distributional factors in the input, related to the frequencies of the constructions reported in Table 1. First, the DO frame is overall much more likely than the PD frame, causing generalization in the PD-only and Alternating conditions to lean more to that frame. Second, fully $1/3$ of the uses of the PD frame in the corpus are with verbs that alternate (i.e., 1% of the corpus are PD frames of alternating PD-DO verbs, out of a total of 3% of the corpus being PD frames), while only $1/8$ of the uses of the DO frame are with alternating rather than non-alternating verbs. Recall that our classes encode the distribution (roughly relative frequency) of the verbs in the class occurring across the different constructions. This means that in our class-based predictions, greater weight will be given to constructions with DO when observing a PD frame than to constructions with PD when observing a DO frame. These results underline the importance of using naturalistic input and considering the impact of various distributional factors on generalization of verb knowledge.

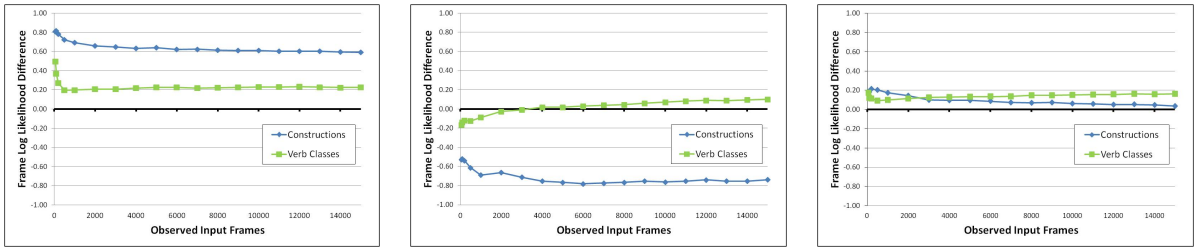
In contrast to the construction-based results, our class-based results conform with the experimental findings of Wonnacott et al. (2008), who show that adult (artificial) language learners robustly generalize a newly-learned verb observed in a single syntactic form by producing it in the alternating syntactic form under certain language conditions. Moreover, we show similar distributional effects to theirs – the overall frequency of the syntactic patterns, as well as the distribution of verbs across those patterns – in the level of preference for one form over another, within the context of our naturalistic data with multiple verbs, constructions, and alternations. These results show that the verb classes in the model are able to capture useful abstract knowledge that is key to understanding the human ability to make high-level generalizations across verb alternations.

5.2 Development of Generalizations

Next, we present the results of our model evaluated throughout the course of training in order to understand the developmental pattern of generalization. We perform the same construction-based or class-based prediction tasks (the likelihoods of a DO and PD frame), following the same input conditions (a novel verb with two DO frames, two PD frames, or one of each) at given points during the $15,000$ inputs. As above, we present the difference in the log-likelihood values of the DO and the PD frames in order to focus on the relative likelihoods of the two frames within each condition of construction-based or class-based predictions.

Figure 3(a) presents the results for the DO-only test scenario. As in Section 5.1, for both construction-based and class-based predictions there is a higher likelihood for the DO frame throughout the course of training. In contrast, the incremental results for the PD-only test scenario, in Figure 3(b), display a developing level of generalization throughout the training stage for the class-based predictions. While the construction-based predictions reflect a much higher likelihood for the PD frame, the results from the verb classes are in favor of the PD frame only initially; after training on 5000 input frames, the likelihood of the DO frame becomes higher for this test scenario. These results indicate that using construction knowledge alone does not enable generalization from the PD frame to the DO frame; in contrast, the verb class knowledge enables the gradual acquisition of generalization ability over the course of training.

Finally, Figure 3(c) presents the results for the Alternating test scenario for the two types of predictions. As in Section 5.1, both construction-based and class-based predictions have a small preference for the DO frame. In the construction-based predictions, this preference lessens over time to where the likelihoods for DO and PD are almost equal, while the class-based predictions stay relatively constant in their preference for the DO frame. In some ways the construction-based predictions are more expected in response to an apparently alternating verb; however, the class-based predictions show a higher degree of generalization, responding to the higher frequency of the



(a) DO only (b) PD only (c) Alternating

Figure 3: Difference of log-likelihood values of the DO and PD frames over the course of training for the constructions and the verb classes for each of the 3 test scenarios. Values above zero denote a higher likelihood for the DO frame, and values below zero denote a higher likelihood for the PD frame.

DO frame and the higher association of PD frames with DO alternates. These results again emphasize the importance of further exploring the role of distributional factors on generalization of verb knowledge in children.

The developmental results presented here are in line with the suggestions of Tomasello (2003) that the productions of younger children follow observed patterns in the input, and only later reflect robust generalizations of their knowledge across verbs. Conwell and Demuth (2007) for example, found evidence of generalization across verb alternations in 3-year-old children, but their production of unobserved forms for a novel verb was very sensitive to the precise context of the experiment and the distributional patterns across the novel verbs. In accord with these observations, the developmental trajectories in our model show that our class-based predictions increase in their degree of generalization over time, and are sensitive to various distributional factors in the input, such as the overall expectation for a frame and the expectation that a verb will alternate.

6 Discussion

We present a novel computational model that probabilistically learns two levels of abstractions over individual verb usages: constructions that are clusters of similar verb usages, and classes of verbs with similar distributional behaviour across the constructions. Specifically, we extend the model of AS08 by incrementally learning token-based verb classes that generalize over the construction knowledge level. In contrast to the models of Parisien and Stevenson and Perfors et al., our model is incremental, and hence enables the analysis of the monotonically developing classes to show the relation to the development of generalization ability in human learners.

We analyze how generalization is supported by each level of learning in our model: constructions and verb classes. Our results confirm (cf. Parisien and Stevenson, 2010) that a higher-level knowledge of the verb classes is required to replicate the observed patterns of generalization, such as producing a novel verb *gorp* in the in the prepositional dative pattern after hearing it in the double object dative pattern. In addition, our analysis of the incrementally developing verb classes shows that the generalization knowledge gradually emerges over time, similar to what is observed in children.

The flexibility of input representation of our model enables us to further explore the properties of the input in learning abstract knowledge, following psycholinguistic studies. Our results replicate the findings of Wonnacott et al. on the role of the distributional properties over the alternating syntactic forms, but in naturalistic settings of many constructions. In future, we plan to extend this analysis by manipulating the distributions of our input data to replicate the exact settings of the artificial language used by Wonnacott et al.. Moreover, in this study, we followed the settings of previous computational and psycholinguistic studies that focused on the syntactic properties of the input (Perfors et al., 2010; Parisien and Stevenson, 2010; Wonnacott et al., 2008; Conwell and Demuth, 2007). However, we can further our analysis by incorporating semantic features in the input to study syntactic bootstrapping effects (Scott and Fisher, 2009) as well as the role of semantic properties in constraining the generalizations across the alternating forms.

Acknowledgments

The authors would like to thank Afra Alishahi for providing us with the code and data for her model, and to Chris Parisien for sharing his data with us.

References

Afra Alishahi and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science*, 32(5):789–834.

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. 2013. Acquisition of desires before beliefs: A computational investigation. In *Proceedings of CoNLL-2013*.

Lois Bloom, Lois Hood, and Patsy Lightbown. 1974. Imitation in language development: If, when, and why. *Cognitive Psychology*, 6(3):380–420.

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

Erin Conwell and Katherine Demuth. 2007. Early syntactic productivity: Evidence from dative shift. *Cognition*, 103(2):163–179.

Cynthia Fisher. 2002. The role of abstract syntactic knowledge in language acquisition: A reply to Tomasello. *Cognition*, 82(3):259–278.

A. Kuczaj, Stan. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600.

B. Levin. 1993. *English verb classes and alternations: A preliminary investigation*, volume 348. University of Chicago press Chicago, IL.

Elena Lieven, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507.

B. MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*, volume 2. Psychology Press.

P. Merlo and S. Stevenson. 2000. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Christopher Parisien and Suzanne Stevenson. 2010. Learning verb alternations in a usage-based bayesian model. In *Proceedings of the 32nd annual meeting of the Cognitive Science Society*.

Amy Perfors, Joshua B. Tenenbaum, and Elizabeth Wonnacott. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37(03):607–642.

Jacqueline Sachs. 1983. Talking about the There and Then: The emergence of displaced reference in parent–child discourse. *Children's language*, 4.

Sabine Schulte im Walde and Chris Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Philadelphia, PA.

Rose M Scott and Cynthia Fisher. 2009. Two-year-olds use distributional cues to interpret transitivity-alternating verbs. *Language and cognitive processes*, 24(6):777–803.

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

Michael Tomasello. 2003. Constructing a language: A usage-based theory of language acquisition.

Michael Tomasello and Kirsten Abbot-Smith. 2002. A tale of two theories: Response to Fisher. *Cognition*, 83(2):207–214.

Elizabeth Wonnacott, Elissa L Newport, and Michael K Tanenhaus. 2008. Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive psychology*, 56(3):165–209.