Comparing Computational Cognitive Models of Generalization in a Language Acquisition Task

Libby Barak, Adele E. Goldberg,
Psychology Department
Princeton University
Princeton, NJ, USA
{lbarak,adele}@princeton.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
Toronto, Canada
suzanne@cs.toronto.edu

Abstract

Natural language acquisition relies on appropriate generalization: the ability to produce novel sentences, while learning to restrict productions to acceptable forms in the language. Psycholinguists have proposed various properties that might play a role in guiding appropriate generalizations, looking at learning of verb alternations as a testbed. Several computational cognitive models have explored aspects of this phenomenon, but their results are hard to compare given the high variability in the linguistic properties represented in their input. In this paper, we directly compare two recent approaches, a Bayesian model and a connectionist model, in their ability to replicate human judgments of appropriate generalizations. We find that the Bayesian model more accurately mimics the judgments due to its richer learning mechanism that can exploit distributional properties of the input in a manner consistent with human behaviour.

1 Introduction

Native speakers of a language are mostly able to generalize appropriately beyond the observed data while avoiding overgeneralizations. A testbed area for studying generalization behavior in language acquisition is verb alternations – i.e., learning the patterns of acceptability of alternative constructions for expressing similar meanings. For example, English speakers readily use a new verb like *text* in both the double-object (DO) construction ("text me the details") and the prepositional-dative (PD) ("text the details to me") – an instance of the dative alternation. However, speakers avoid overgeneralizing the DO construction to verbs such as *explain* that resist its use ("?explain me the details"), even though they occur with analogous arguments in the PD alternative ("explain the details to me"). Psycholinguistic studies have focused on the possible properties of natural language that enable such generalization while constraining it to acceptable forms.

Initially, children are linguistically conservative: they generally use verbs in constructions that are very close to exemplars in the input (Lieven et al., 1997; Akhtar, 1999; Tomasello, 2003; Boyd and Goldberg, 2009). Children reach adult-like competence by gradually forming more general associations of constructions to meaning that allow them to extend verb usages to unwitnessed forms. Much work has emphasized the role of verb classes that capture the regularities across semantically-similar verbs, enabling appropriate generalization (e.g., Pinker, 1989; Fisher, 1999; Levin, 1993; Ambridge et al., 2008). Usage-based approaches have argued that such class-based behaviour can arise in learning through the clustering of observed usages that share semantic and syntactic properties (e.g., Bybee, 2010; Tomasello, 2003; Goldberg, 2006).

A number of studies also reveal that the statistical properties of the language play a central role in limiting generalization (e.g., Bresnan and Ford, 2010; Ambridge et al., 2012, 2014). Individual verbs often show statistical biases that favor their appearance in one construction over another (Ford et al., 1982; MacDonald et al., 1994; Garnsey et al., 1997; Trueswell et al., 1993; Losiewicz, 1992; Gahl and Garnsey, 2004). For example, while both *give* and *push* can occur in either DO or PD constructions,

give strongly favors the DO construction ("give me the box"), while *push* strongly favors the PD ("push the box to me") (Wasow, 2002). Generally, the more frequent a verb is overall, the less likely speakers are to extend it to an unobserved construction (Braine and Brooks, 1995). In addition, when a verb repeatedly occurs in one construction when an alternative construction could have been appropriate, speakers appear to learn that the verb is inappropriate in the alternative, regardless of its overall frequency (Goldberg, 2011).

Given these observations, it has been argued that both the semantic and statistical properties of a verb underlie its degree of acceptability in alternating constructions (e.g., Braine and Brooks, 1995; Theakston, 2004; Ambridge et al., 2014). Recently, Ambridge and Blything (2015) propose a computational model designed to study the role of verb semantics and frequency in the acquisition of the dative alternation. However, they only evaluate their model preferences for one of the two constructions, which does not provide a full picture of the alternation behaviour; moreover, they incorporate certain assumptions about the input that may not match the properties of naturalistic data.

In this paper, we compare the model of Ambridge and Blything (2015) to the Bayesian model of Barak et al. (2014) that offers a general framework of verb construction learning. We replicate the approach taken in Ambridge and Blything (2015) in order to provide appropriate comparisons, but we also extend the experimental settings and analysis to enable a more fulsome evaluation, on data with more naturalistic statistical properties. Our results show that the Bayesian model provides a better fit to the psycholinguistic data, which we suggest is due to its richer learning mechanism: its two-level clustering approach can exploit distributional properties of the input in a manner consistent with human generalization behaviour.

2 Related Work

Acquisition of the dative alternation – use of the DO and PD constructions with analogous semantic arguments – has been studied in several computational cognitive models because it illustrates how people learn to appropriately generalize linguistic constructions in the face of complex, interacting factors. As noted by Ambridge et al. (2014), such models should capture influences of the verb such as its semantic properties, its overall frequency, and its frequency in various constructions.

A focus of computational models has been to show under what conditions a learner generalizes to the DO construction having observed a verb in the PD, and vice versa. For example, the hierarchical Bayesian models of Perfors et al. (2010) and Parisien and Stevenson (2010) show the ability to generalize from one construction to the other. However, both models are limited in their semantic representations. Perfors et al. (2010) use semantic properties that directly (albeit noisily) encode the knowledge of the alternating and non-alternating (DO-only or PD-only) classes. The model of Parisien and Stevenson (2010) addresses this limitation by learning alternation classes from the data (including the dative), but it uses only syntactic slot features that can be gleaned automatically from a corpus. In addition, both models use batch processing, failing to address how learning to generalize across an alternation might be achieved incrementally.

Alishahi and Stevenson (2008) presents an incremental Bayesian model shown to capture various aspects of verb argument structure acquisition (Alishahi and Pyykkoïen, 2011; Barak et al., 2012, 2013b; Matusevych et al., 2016), but the model is unable to mimic alternation learning behaviour. Barak et al. (2014) extends this construction-learning model to incrementally learn both constructions and classes of alternating verbs, and show the role of the classes in learning the dative. However, like Parisien and Stevenson (2010), the input to the model in this study is limited to syntactic properties, not allowing for a full analysis of the relevant factors that influence acquisition of alternations.

Ambridge and Blything (2015) propose the first computational model of this phenomenon to include a rich representation of the verb/construction semantics, drawn from human judgments. In evaluation, however, they only report the ability of the model to predict the DO usage (i.e., only one pair of the alternation), which does not give the full picture of the alternation behaviour. Moreover, their assumptions about the nature of the input – including the use of raw vs. log frequencies and the treatment of

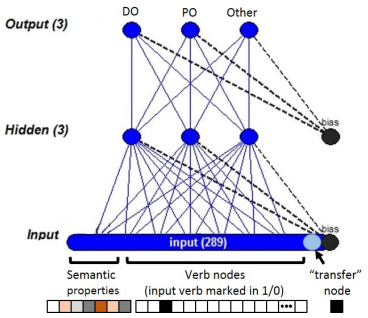


Figure 1: A visual representation of the feed-forward network used by the AB model. (The figure is adapted from output of the OXlearn package of Ruh and Westermann (2009).) The input nodes correspond to the semantic properties of the verbs, the verb lexemes, and a "transfer" node (explained in the text). The output nodes correspond to the target constructions.

non-dative construction usages – differ from earlier models, making it difficult to compare the results.

In this paper, we compare the models of Ambridge and Blything (2015) and Barak et al. (2014), using the same input settings for each, so that, for the first time, two computational models of this generalization phenomenon can be directly compared. Moreover, in contrast to Ambridge and Blything (2015) and in line with the other studies mentioned above, we evaluate the ability of the models to generate both the DO and the PD alternates, on a per verb basis, in order to more accurately assess the fit to human judgments.

3 The Computational Models

In this section, we give an overview of the connectionist model of Ambridge and Blything (2015), hereafter the AB model, and the Bayesian model of Barak et al. (2014), hereafter the BFS model, followed by a comparison of their relevant properties.

3.1 Overview of the Connectionist Model

The AB connectionist model of Ambridge and Blything (2015) aims to predict the preference of a verb for each of three target constructions, on the basis of verb semantics and the observed distribution of verbs in those constructions in the input. Figure 1 provides an illustration of the 3-layer feed-forward network, trained using backpropagation. Each input to the model consists of lexical and semantic features of a verb and its usage. The target output is a 1-hot pattern across output nodes, each of which represents the use of the verb in the associated construction. The possible constructions are DO, PD, or *other*, representing all other constructions the verb appears in. Training presents the slate of input features with the appropriate output node activated representing the construction the verb appears in. In a full sweep of training, the model observes all verbs in proportion to their frequency in the input; for each verb, the proportion of training trials with 1 in each of the output nodes corresponds to the frequency of the verb in each of those constructions. During testing, only the input nodes are activated (corresponding to a verb and its semantics), and the activation of output nodes reveals the learned proportional activation rate corresponding to the degree of verb bias toward either the DO or the PD (or *other*).

The structure of the AB model encodes some assumptions regarding the information and learning mechanisms available to the learner. The model incorporates awareness of individual verbs by having a node per verb in the input to distinguish the usage of each verb and its accompanying features. Each verb is also represented by a vector of semantic features that capture properties relevant to its meaning when used in one of the two dative constructions (based on elicited human judgments from Ambridge et al., 2014). The "transfer" input node encodes the ability to distinguish the semantic properties of the dative constructions from other constructions: i.e., this node is set to 1 for a DO or PD usage, and to 0 otherwise. Representing the construction of the input usage (DO, PD, or *other*) on the output nodes reflects the formalization of the learning as an association of semantic and lexical features with a syntactic pattern, and the knowledge of the model is demonstrated by activating the construction output nodes in response to a lexical/semantic input.

3.2 Overview of the Bayesian Model

The BFS of model Barak et al. (2014) is a Bayesian clustering model that simultaneously and incrementally learns both constructions and verb classes in a two-level design; see Figure 2 for an illustration of each level. In learning, the model processes an input sequence of verb usages, represented as collections of semantic and syntactic features, one usage at a time. The first step of processing each input aims to

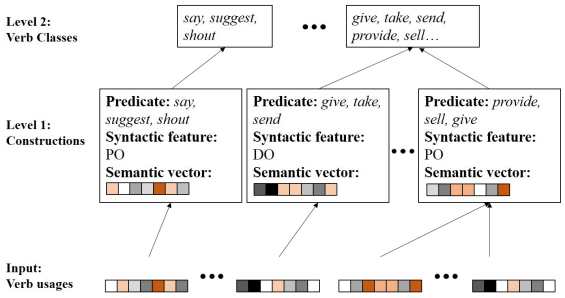


Figure 2: A visual representation of the the Bayesian model, with sample input features for verb usages, construction level, and verb class level.

find the best cluster at level one as:

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

where F_i is the set of features for input i, and k ranges over all existing clusters and a new one. The number of possible clusters is not set in advanced, and thus at any step the best choice may be to start a new cluster (of size 1) with this input.

Using Bayes rule:

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

The prior probability of a cluster $P(k)$ is proportional to the number of verb usages clustered to k so far, thus assigning a higher prior to larger clusters. The likelihood $P(F_i|k)$ is estimated based on the match of feature values in the current verb usage to those aggregated in the cluster, where the quality of the match depends on the frequency and vector similarity of the two sets of features.

The clusters at this level correspond to constructions of the language – i.e., probabilistic associations of form and meaning. For example, a cluster emerges from semantically-similar verbs like *tell* and *ask*, in a particular syntax, such as the DO. Creating a new cluster – forming a new construction – depends on both the likelihood and the prior. Early on, the $P(F_i|k)$ term has more influence and differences in feature values between a new usage and existing clusters will often trigger a new cluster. Later, the model will favour adding a new input to an existing cluster – even if it makes it more heterogeneous – because the $P(k)$ term prefers larger clusters as the number of observed inputs increases. This mechanism mimics human language learning behavior of moving from more verb-specific constructions to more general constructions (Tomasello, 2003).

Each verb can occur in several clusters in the first level based on its association with various semantic and syntactic features. For instance, the alternating verb *give* can occur in one cluster associating a transfer meaning with PD syntax and in a second cluster associating a transfer meaning with DO syntax. To capture the common behaviour of such alternating verbs – where verbs with similar meanings occur across the same set of clusters – the model represents the similarity in distributional properties of the verbs in a second level of representation, which captures such verb class behaviours.

Formally, after each clustering decision in the first level, the model calculates the current frequency distribution of the input verb over all level-one clusters. This distribution vector is used as input for the second level: the model measures the similarity of this vector to the weighted average distribution represented by each second-level cluster, adding the input verb's distribution to the most similar one:

$$\text{BestClass}(d_{v_t}) = \underset{c \in Classes}{\text{argmax}}\ (1 - D_{\text{JS}}(d_c \| d_{v_t})) \quad (3)$$

where d_{v_t} is the current distribution of the verb v over the clusters (i.e., at time t), c ranges over all the classes in the second level, d_c is the weighted average of c given the distributions of its member verb tokens, and D_{JS} is the Jensen–Shannon divergence. As in the first level, the model may create a new cluster in the second level if none of the existing clusters is similar enough to d_{v_t}.

The resulting second-level clusters capture verb class behavior by grouping verbs that share a pattern of usages across constructions, e.g., alternating verbs that occur with the DO and PD syntax. These clusters encode a snapshot of the distribution of each verb each time it occurs in the input, reflecting the need of the language learner to incrementally update their knowledge of the distributional behavior of the verb across constructions.

3.3 Comparison of the Models

Both models capture the semantic and statistical properties of language proposed as possible factors in the ability to learn an alternation appropri-

ately – i.e., to generalize to new uses but not overgeneralize to inappropriate verbs. Semantic influences are reflected in the use of meaning features, and each model incorporates the key idea behind statistical preemption (Goldberg, 1995), namely that semantically-appropriate constructions compete with one another. The statistical effects of overall verb frequency and of frequency of the verb-in-construction are captured by inputting each verb in proportion to its frequency with each construction.

The models have a crucial difference in how they reflect the influence of the various features in learning alternations. As a feed-forward network, the AB model learns the weight of the semantic features given the entire set of input, and uses these weightings to shape the prediction of a verb's preference for each of the syntactic constructions (represented by the target output nodes). The BFS model does not explicitly weight features, but the influence of a feature is determined by its local context within a cluster. For example, if the value of a feature has high frequency in a cluster – e.g., the cluster records usages with only the DO syntax – the predictions based on this cluster would strongly prefer matching usages based on this feature value; a less-frequent feature would have less influence on this cluster's predictions, but could have more influence in a cluster where it is more represented. This property, along with the representation of an open-ended set of constructions (level one clusters) and verb classes (level two clusters), enables the model to capture rich interactions among the lexical, semantic, and syntactic features. We evaluate the role of these differences in the fit of each model to the task.

4 Experimental Setup

4.1 Input and Training

We base the learning and evaluation on the 281 distinct verbs used in Ambridge and Blything (2015), which had been determined to occur in the double object (DO) and/or the preposition-dative (PD) (Pinker, 1989; Levin, 1993). Following Ambridge and Blything (2015), we consider a third (artificial) construction labeled as *other* that corresponds to *all* non-DO and non-PD usages of a verb. The models are trained on usages of the verbs in proportion to their frequencies in the British National Corpus (BNC) (Leech, 1992). Table 1 summarizes the per-construction frequency data for the verbs.[1] Note that 98 of the verbs can occur in the DO and/or PD but have no such occurrences in the BNC; these verbs unseen in the dative are important for judging the appropriate generalization behavior of the models.

		Raw freq			Log freq		
	#Verbs	DO	PD	*other*	DO	PD	*other*
PD	101	0	93	9964	0	3	5
DO	7	49	0	877	2	0	4
Alt	75	325	1144	13332	3	3	7
Uns	98	0	0	716	0	0	4

Table 1: Frequency data for the dative verbs in the BNC for non-alternating **PD**-only and **DO**-only verbs, **ALT**ernating verbs, and **UNS**een dative-taking verbs that do not occur with the dative constructions in the BNC.

The input to the models include: the lexeme of the verb, the semantic features of the verb, a "transfer" feature marking the common meaning of the dative constructions, and (in training only) a syntactic feature. The syntactic feature indicates whether a verb is used with the DO, PD, or *other* construction; in the AB model, this is given as the target output node in training. The verb semantic features are those used in Ambridge and Blything (2015). These vectors are based on the ratings of each verb on 18 meaning properties relevant to use of the verb in the dative (e.g., "The verb specifies the means of transfer" Ambridge et al., 2014), subject to Principal Component Analysis by Ambridge and Blything (2015), yielding a vector of 7 dimensions. The transfer feature is 1 for a verb usage in one of the two dative constructions, and 0 for the *other* construction, to indicate the shared transfer meaning conveyed by the DO and the PD. The input to each model is generated automatically to correspond to the BNC frequencies of each verb in each of the constructions.

It should be noted that, while we adopt the semantic features of Ambridge and Blything (2015), they reflect the meaning within the two dative constructions and may be less applicable to the *other* construction. In addition, we found that there are alternating and non-alternating verbs that have very similar semantic vectors, indicating that these fea-

[1] The full list of verbs and their frequencies can be found in Ambridge and Blything (2015).

tures may not sufficiently distinguish the alternation behaviours.

The models are trained with sufficient input to converge on stable behavior. We follow Ambridge and Blything (2015) in training and testing the AB model using the OXlearn MATLAB package (Ruh and Westermann, 2009); the input is generated using a random seed, in random input order without replacements, and the model is trained with a learning rate of 0.01 for 1K sweeps for log frequencies; 100K sweeps for raw frequencies. We train the BFS model using the input generation method described by Barak et al. (2014), with the features as above. The model is trained on 5K input verb usages (in proportion to their frequencies in the constructions).

4.2 Evaluation of the Models

As in Ambridge and Blything (2015), to test the model preferences for the DO or PD, the models are presented with an input consisting of a verb lexeme, its semantic features, and the transfer feature set to 1 (i.e., this is a "transfer" semantics suitable for a dative construction). For the AB model, we measure preferences for each construction as the activation rate of each of the corresponding output nodes, as in Ambridge and Blything (2015). In the BFS model, the preference for each construction is measured as its likelihood over the learned clusters given the verb and its semantic features. Formally, the prediction in the Bayesian model is:

$$P(s|F_{\text{test}}) = \sum_{k \in Clusters} P(s|k)P(k|F_{\text{test}}) \quad (4)$$

where s is the predicted syntactic construction (DO or PD) and F_{test} is the set of test features representing a verb v and its corresponding semantic features. $P(s|k)$ is the probability of the syntactic pattern feature having the value s in cluster k, calculated as the proportional occurrences of s in k. $P(k|F_{\text{test}})$ is the probability of cluster k given test features F_{test}, calculated as in Eqn. (2). Following Barak et al. (2014), we calculate $P(k|F_{\text{test}})$ in two ways, using just the constructions (level one) or both the classes (level two) and the constructions, to see whether verb class knowledge improves performance. Using solely the construction level, the probability of k reflects the frequency with which usages of verb v occur in cluster k. Using the verb class level in addition, the distribution of the verb over classes in the second level is combined with the distribution of those classes over the constructions in level one, to get the likelihood of k.

These model preferences of the verbs for a dative construction are compared, using Pearson correlation, to the DO/PD acceptability judgment data collected from adult participants by Ambridge et al. (2014). Note that Ambridge and Blything (2015) only evaluate their model's preferences for verbs to take the DO construction. To fully understand the preference and generalization patterns, we also analyze the results for the PD preference. Even more importantly, we calculate the *difference* between the preferences for the DO and the PD constructions *per verb*, and compare these to analogous scores for the human data, as suggested by Ambridge et al. (2014). The DO$-$PD difference scores, which we will refer to as the **verb bias score**, are crucial because, as in the human data, it is these scores that accurately capture a learner's relative preference for a construction *given a particular verb*.

5 Experiments and Analysis of Results

We examine the ability of each model to match the dative construction preferences of human judgments, as described just above, under two different experimental scenarios. In Section 5.1, we follow the experimental settings of Ambridge and Blything (2015). We replicate their results on the AB model showing correlation with human DO preferences, but find that only the BFS model achieves a significant correlation with the crucial verb bias score that appropriately assesses per-verb preference. We adjust the experimental settings in Section 5.2 to use more naturalistic input data – by training in proportion to raw frequencies and excluding the artificial *other* construction – achieving an improvement in the verb bias score for both models.

5.1 Exp 1: Log Freq Input; 3 Constructions

Results. We first evaluated the models under the experimental conditions of Ambridge and Blything (2015), providing input corresponding to the verbs in 3 constructions (DO, PD, and *other*), in proportion to their log frequencies; see Table 2. We replicate the positive correlation of the AB model over

	AB (Connectionist)	BFS (Bayesian)	
		Level 1	Level 2
DO	**0.54**	0.24	0.29
PD	0.39	0.30	**0.50**
DO-PD	[-0.02]	0.48	**0.53**

Table 2: Pearson correlation values between human and model preferences for each construction and the verb-bias score (DO−PD); training on log frequencies and 3 constructions. All correlations significant with p-value < 0.001, except the one value in square brackets. Best result for each row is marked in boldface.

the ratings for the DO construction found in Ambridge and Blything (2015). In addition, our analysis shows that the AB model produces a significant positive correlation with the PD acceptability rating. However, the AB model has no correlation with the verb bias score. Although the model ranks the separate verb preferences for DO and PD similarly to humans, the model does not produce the same relative preference *for individual verbs*. For example, the human data rank *give* with high acceptability in both the DO and the PD, with a higher value for the DO construction. Although the AB model has a high preference for both constructions for *give* (compared with other verbs), the model erroneously prefers *give* in the PD construction.

The BFS model also produces preferences for verbs in each construction that have a significant positive correlation with human judgments. While the AB model shows better correlation with the DO judgments, the BFS model correlates more strongly with the PD judgments. Importantly, in contrast to the AB model, the verb bias score of the BFS model also significantly correlates with the judgment data. That is, the BFS model provides a better prediction of the preference per verb, which is key to producing a verb in the appropriate syntax.

Analysis. We can explain these results by looking more closely at the properties of the input and the differences in the learning mechanisms of each model. Following Ambridge and Blything (2015), the input presents an artificial *other* construction in proportion to the frequency of the verbs with all non-dative constructions. The very high frequency of this single artificial construction (see *other* in Table 1) results in higher predictions of it for any of the verbs, even though the "transfer" feature in test inputs has a value intended to signal one of the dative constructions. As a result, the preferences for the dative constructions in both models have a very small range of values, showing relatively small differences.

The BFS model is also affected by the relatively compressed semantic space of the input, which is exacerbated by the use of log frequencies to guide the input. As noted earlier, we found that the semantic features of alternating verbs can be highly similar to non-alternating verbs – e.g., *give* (alternating) and *pull* (PO-only) have similar semantic vectors. With such input, the model cannot form sufficiently distinct first-level clusters based on the semantics, particularly when the data is presented with such a flat distribution (note the small differences in log frequencies in Table 1). Visual inspection reveals that these clusters in the model largely form around syntactic constructions, with mixed semantic properties. Despite this, the first-level clusters capture a strong enough association between individual verbs and their constructions to yield a good correlation of the verb bias score with human judgments, and drawing on the second-level (verb-class) clusters improves the results.

Conclusions. The use of an artificial high-frequency non-dative construction (*other*), and the use of log frequencies, seem to mask the influence of the semantic and syntactic properties on learning the verb-bias for each verb. Previous psycholinguistic data and computational models have found that a skewed naturalistic distribution of the input is helpful in learning constructions, due to the high-frequency verbs establishing appropriate construction-meaning associations (Casenhiser and Goldberg, 2005; Borovsky and Elman, 2006; Barak et al., 2013b; Matusevych et al., 2014). To allow a more direct analysis of the role of statistical and semantic properties in learning and generalizing the dative, we adjust the input to the models in the next section.

5.2 Exp 2: Raw Freq Input; 2 Constructions

Results. Here we perform the same type of experiments, but using input in proportion to the raw frequencies of the verbs (instead of log frequencies) over occurrences only in the two dative constructions (with no *other* construction). Since 98 of the 281 verbs do not occur with either dative construc-

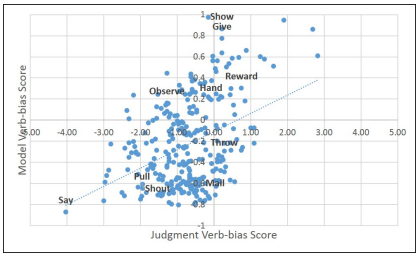
	AB (Connectionist)	BFS (Bayesian)	
		Level 1	Level 2
DO	[0.06]	0.23	**0.25**
PD	0.33	**0.38**	0.32
DO-PD	0.39	0.53	**0.59**

Table 3: Pearson correlation values between human and model preferences for each construction and the verb-bias score; training on raw frequencies and 2 constructions. All correlations significant with p-value < 0.001, except the one value in square brackets. Best result for each row is marked in boldface.

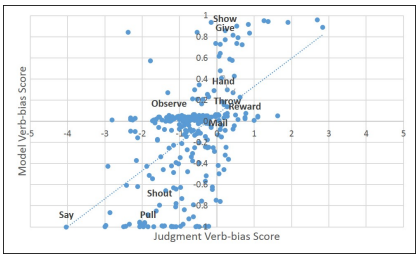
tion in the BNC, this also allows us to more stringently test the generalization ability of the models, by considering their behavior when ˜$1/3$ of the verbs are unseen in training.

Table 3 presents the correlation results for the two models' preferences for each construction and the verb bias score; we also show the correlation plots for the verb bias score in Figure 3. The AB model does not correlate with the judgments for the DO. However, the model produces significant positive correlations with the PD judgments and with the verb bias score. The BFS model, on the other hand, achieves significant positive correlations on all measures, by both levels. As in the earlier experiments, the best correlation with the verb bias score is produced by the second level of the BFS model, as Figure 3 demonstrates.

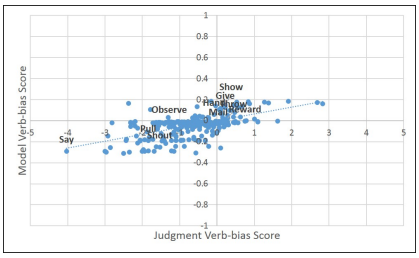
Analysis. As shown by Barak et al. (2013b), the Bayesian model is better at learning the distribution pattern of each verb class given a skewed distribution, as in the raw frequencies here. The model learns an association of each construction to the frequently-observed meaning of high-frequency verbs. For example, the semantics of the DO is most strongly influenced by the semantics of its most frequently occurring instance: *give*. The accuracy of preference judgments benefits from the entrenchment of the relevant meaning with the construction. This supports appropriate generalization – e.g., because *reward* is semantically similar to *give*, it has a good fit to the human preference judgments even though it is unseen with the dative (see Figure 3). But the same factor can serve to limit generalization – e.g., because the unseen verb *mail* is semantic *dissimilar* to a frequent PD-only verb like *pull*, its preference for the PD syntax is limited, giving it a good match to human judgments by preventing its overgeneralization (see Figure 3).



(a) AB model



(b) BFS model - construction level



(c) BFS model - verb class level

Figure 3: Correlation of the dative verbs with the verb bias score of each model in Exp. 2: (a) the AB model ($r = 0.39$), (b) the first level of the BFS model ($r = 0.53$), and (c) the second level of the BFS model ($r = 0.59$).

The AB model can also take advantage of high frequency verbs biasing the preference toward the frequently observed association. However, the semantic similarity across verbs within alternating or non-alternating classes is less effective in this model. The representation of the lexemes as 281 nodes in the input (compared to less than a dozen other nodes) make the learning more verb specific, reducing the ability of the model to generalize to the unattested verbs.

Conclusions. The success of the BFS model, and especially the results using both constructions and classes, point to the role of probabilistic constructions and verb classes in generalizing existing knowledge while avoiding overgeneralizations.

Moreover, the use of a skewed distribution reveals the role of the high verb-in-construction frequency in guiding the association of construction and meaning (see Ambridge et al., 2014, for discussion). Yet both models investigated here would benefit from a richer semantic representation that better captures the distinctive properties of verbs across various constructions.

6 Discussion

This paper presents a comparative analysis of two computational cognitive models on the sample task of learning the dative alternation. This study enables an evaluation of the psycholinguistic plausibility of each model for the given task when facing identical input and experimental settings. Adopting the semantic representation of Ambridge and Blything (2015), our input incorporates both semantic and syntactic properties over a large number of verbs. By providing the first direct comparison between two existing models of this phenomenon, we are the first to demonstrate the complex interaction of various linguistic properties in the input, and how rich learning mechanisms are required in order to achieve generalizations compatible with human judgments in this area. Moreover, comparison of learning mechanisms and of input properties can inform CL/NLP more generally by shedding light on potential factors in achieving humanlike behaviours.

We find that the Bayesian model of BFS significantly correlates with human judgments on the 3 key evaluation measures. Importantly, this model outperforms the connectionist model of AB in the correlation with the verb-bias score (the per-verb difference between DO and PD preference), which points to its advantage in choosing the more appropriate construction per verb. We argue that the fit of the model relies on a rich learning mechanism that exploits distributional properties of naturalistic input.

The AB model has a streamlined design to support learning a particular semantic-syntactic association underlying the dative alternation. While the BFS model is richer computationally, its properties were motivated in earlier work explaining many human behaviours. When we consider more natural input, the simple input[semantics]–output[syntax] association mechanism of the AB model is unable to capture the necessary interactions among the verb semantic properties, the syntactic usages, and their patterns across different types of verbs. By contrast, the two-level design of the BFS model captures these interactions. The first level learns the verb-semantics-syntax associations as clusters of similar configurations of those features. The second level captures the commonalities of behaviour of sets of verbs by forming classes of verbs that have similar distributional patterns over the first-level clusters. We also observe that the replication of adult-like language competence relies on several naturalistic properties of the input: skewed distribution, and a rich semantic representation combined with syntactic information. The skewed input enables the formation of clusters representing more entrenched associations, which are biased towards high-frequency verbs associated with certain semantic and syntactic features.

Given the role of these linguistic properties, the results here call for additional analysis and development of the input to computational cognitive models. The predictions may be improved given more realistic syntactic and semantic information about the verb usages. On the syntax side, the input should reflect the distribution of verbs across more syntactic constructions, as statistical patterns over such usages can indirectly indicate aspects of a verb's semantics (cf. Barak et al., 2013a). In the future, we aim to analyze the role of fuller syntactic distributions in restricting overgeneralization patterns. Moreover, the semantic annotations used here replicate the settings originally tested for the AB model, which correspond to the verb as used in the relevant constructions. This contrasts with typical automated extractions of verb-meaning representation (e.g., word2vec, Mikolov et al., 2013), which capture a more general verb meaning across all its usages. In preliminary experiments, we have found an advantage in using word2vec representations in addition to the semantic properties reported here. We aim to further analyze manual and automated methods for semantic feature extraction in future work.

Acknowledgments

We are grateful to Ben Ambridge for helpful discussion of his model and for sharing his data with us.

References

Nameera Akhtar. 1999. Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of child language*, 26(02):339–356.

Afra Alishahi and Pirita Pyykkoñen. 2011. The onset of syntactic bootstrapping in word learning: Evidence from a computational study. In *Proceedings of the 33st Annual Conference of the Cognitive Science Society*.

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

Ben Ambridge and Ryan P Blything. 2015. A connectionist model of the retreat from verb argument structure overgeneralization. *Journal of child language*, pages 1–32.

Ben Ambridge, Julian M Pine, Caroline F Rowland, and Franklin Chang. 2012. The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language*, 88(1):45–81.

Ben Ambridge, Julian M Pine, Caroline F Rowland, Daniel Freudenthal, and Franklin Chang. 2014. Avoiding dative overgeneralisation errors: Semantics, statistics or both? *Language, Cognition and Neuroscience*, 29(2):218–243.

Ben Ambridge, Julian M Pine, Caroline F Rowland, and Chris R Young. 2008. The effect of verb semantic class and verb frequency (entrenchment) on childrens and adults graded judgements of argument-structure overgeneralization errors. *Cognition*, 106(1):87–129.

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. 2012. Modeling the acquisition of mental state verbs. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*.

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

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. 2013b. Modeling the emergence of an exemplar verb in construction learning. In *Proceedings of the 35rd Annual Meeting of the Cognitive Science Society*.

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. 2014. Learning verb classes in an incremental model. In *Proceedings of the 5th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2014)*. Association for Computational Linguistics.

Arielle Borovsky and Jeff Elman. 2006. Language input and semantic categories: A relation between cognition and early word learning. *Journal of child language*, 33(04):759–790.

Jeremy K Boyd and Adele E Goldberg. 2009. Input effects within a constructionist framework. *The Modern Language Journal*, 93(3):418–429.

Martin DS Braine and Patricia J Brooks. 1995. Verb argument structure and the problem of avoiding an overgeneral grammar. *Beyond names for things: Young children's acquisition of verbs*, pages 353–376.

Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

David Casenhiser and Adele E. Goldberg. 2005. Fast mapping between a phrasal form and meaning. *Developmental Science*, 8(6):500–508.

Cynthia Fisher. 1999. From form to meaning: A role for structural alignment in the acquisition of language. *Advances in child development and behavior*, 27:1–53.

Marilyn Ford, Joan W Bresnan, and Ronald Kaplan. 1982. A competence-based theory of syntactic closure. *American Journal of Computational Linguistics*, 8(1):49.

Susanne Gahl and Susan M Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, pages 748–775.

Susan M Garnsey, Neal J Pearlmutter, Elizabeth Myers, and Melanie A Lotocky. 1997. The contributions of verb bias and plausibility to the com-

prehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37(1):58–93.

Adele E. Goldberg. 1995. *Constructions, A Construction Grammar Approach to Argument Structure*. {Chicago University Press}.

Adele E Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Adele E Goldberg. 2011. Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics*, 22(1):131–153.

Geoffrey Leech. 1992. 100 million words of english: the british national corpus (BNC). *Language Research*, 28(1):1–13.

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

Elena VM Lieven, Julian M Pine, and Gillian Baldwin. 1997. Lexically-based learning and early grammatical development. *Journal of child language*, 24(01):187–219.

Beth L Losiewicz. 1992. *The effect of frequency on linguistic morphology*. University of Texas.

Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676.

Yevgen Matusevych, Afra Alishahi, and Ad Backus. 2014. Isolating second language learning factors in a computational study of bilingual construction acquisition. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 988–994.

Yevgen Matusevych, Afra Alishahi, and Ad Backus. 2016. The impact of first and second language exposure on learning second language constructions. *Bilingualism: Language and Cognition*, pages 1–22.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

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.

Steven Pinker. 1989. *Learnability and cognition: The acquisition of argument structure.* The MIT Press.

Nicolas Ruh and Gert Westermann. 2009. Oxlearn: A new matlab-based simulation tool for connectionist models. *Behavior research methods*, 41(4):1138–1143.

Anna L Theakston. 2004. The role of entrenchment in childrens and adults performance on grammaticality judgment tasks. *Cognitive Development*, 19(1):15–34.

Michael Tomasello. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

John C Trueswell, Michael K Tanenhaus, and Christopher Kello. 1993. Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3):528.

Thomas Wasow. 2002. *Postverbal behavior*. Stanford Univ Center for the Study.