Linguistic generalization on the basis of function and constraints on the basis of statistical preemption

Florent Perek and Adele E. Goldberg

A key question in language learning is what encourages and what constraints generalization beyond what is witnessed in the input. Experiment 1 exposes participants to two novel word order constructions that differ in terms of their semantics: One construction exclusively describes actions that have a strong effect; the other construction describes actions with a weaker but otherwise similar affect. One group of participants witnessed novel verbs only appearing in one construction or the other, while another group witnessed a minority of verbs alternating between constructions. Subsequent production and judgment results demonstrate that participants in both conditions extended and accepted verbs in whichever construction best described the intended message. Unlike related previous work, this finding is not naturally attributable to prior knowledge of the likely division of labor between verbs and constructions. A second experiment included one verb (out of six) that was witnessed in a single construction to describe both strong *and* weak effects, essentially preempting the use of the other construction. In this case, participants were much more lexically conservative with this verb *and* other verbs, while they nonetheless displayed an appreciation of the distinct semantics of the constructions with new novel verbs. Results indicate that the need to better express an intended message encourages generalization, while statistical preemption constrains generalization by providing evidence that verbs are restricted in their distribution.

Keywords: language acquisition, artificial language learning, novel construction learning, statistical learning, argument structure constructions, generalization

1. Introduction

Learners sometimes generalize beyond their input and produce verbs in novel ways. For example, by the time children are in preschool, they readily extend nonsense verbs that have only been witnessed intransitively (*It meeked*) for use in the transitive construction (*She meeked it*) (e.g., Akhtar 1999; Tomasello 2000), and their comprehension of familiar and novel verbs used in constructions that are new for those verbs begins even earlier (e.g., Naigles 2000; Fisher 2002).

And yet while speakers produce and comprehend language that goes beyond their input, there are certain generalizations that are only rarely made, and are judged to be less than fully acceptable, even though they are easily understood (Bowerman 1990; Pinker 1989; Goldberg 1995). This type of *overgeneralization* is illustrated by the examples in (1) - (3):

(1) ?? The child seems sleeping (Chomsky 1957)
(2) ?? Don't giggle me (Bowerman 1990)
(3) ?? an asleep boy (Boyd & Goldberg 2011)

When and why do speakers generalize beyond their input? And when and why do they not? These questions have puzzled researchers for decades (Ambridge et al. 2012; Baker 1970; Bowerman 1988; Braine 1990; Goldberg 1995; Lakoff 1970; Perek 2015; Pinker 1989). Recent work has found artificial language learning experiments helpful in addressing these questions (e.g., Braine et al. 1990; Brooks et al. 1993; Casenhiser & Goldberg 2005; Culbertson, Legendre, & Smolensky 2012; Fedzechkina, Jaeger & Newport 2012; Gómez& Gerken 2000; Moeser & Bregman 1972; Valian and Coulson 1988; Amato & MacDonald 2010). A typical paradigm involves exposing learners to a miniature language which includes a set of novel phrases or sentences that are paired with interpretations.

In a study that is particularly relevant in the present context, Wonnacott et al. (2008) observed that the overall statistics of an artificial language play a role in whether predicates are extended in new ways. They demonstrated that adult learners who were exposed to a "lexicalist" language in which most verbs appeared in only one construction behaved conservatively, avoiding extending verbs for use in a different construction; on the other hand, learners exposed to a "generalist" language in which the majority of verbs alternated, appearing in both of two constructions, readily assumed that all verbs alternated (see also Perek & Goldberg 2015, Exp. 2; Thothathiri and Rattinger 2016, Exp.1). Wonnacott (2011) is a similar study that replicated the basic findings with children.

In these studies, two distinct formal patterns were assigned the exact same function. But in natural languages, it is hard to find verbs that occur in two constructions which serve exactly the same function; instead the choice between two constructions is typically conditioned by differences in information structure or semantics (e.g., Bolinger 1971; Bresnan 2011; Goldberg 1995; Scott-Phillips, Kirby, & Ritchie 2009). With this in mind, a previous study by the authors (Perek & Goldberg 2015, Exp. 1) presented learners with six nonce verbs that were used in two constructions that differed in terms of word order *and* information structure properties. In particular, one construction was always used with a pronominal undergoer argument ($\text{Pronoun}_{\text{Undergoer}}$ NP_{Agent} V), while the other occurred exclusively with lexical noun phrase arguments in a distinct order (NP_{Agent} $NP_{Undergoer}$ V). Results demonstrated that learners used verbs in ways that went beyond the verb-specific regularities in the input in order to take advantage of the information structure properties of the newly learned constructions. More specifically, when even a minority of the verbs in the input alternated, participants freely used all of the verbs in whichever construction was more appropriate in the given discourse context, ignoring the fact that most of the verbs had been witnessed only in one construction or the other. Even in a lexicalist condition, in which each of the six verbs in the input appeared only in one construction or the other, participants still showed a tendency to generalize beyond their input, although they were also lexically conservative to a lesser extent.

Similarly, Thothathiri and Rattinger (2016) exposed adult participants to a mini-artificial language in order to determine whether learners tended to generalize on the basis of verb-specific information or on the basis of the functions of the constructions. When verbs were witnessed in one of two transitive constructions that had no discernable difference in function, participants were lexically conservative. When the two constructions did have distinct functions, Thothathiri and Rattinger (2016) found a strong tendency to generalize on the basis of the functions constructions, using verbs in whichever construction better captured the intended message. In their experiment, one construction had Verb-Agent-Patient order and included an additional, final nominal that was interpreted as an instrument, and the other construction had V-Patient-Agent order and included a final nominal that was interpreted as a modifier (something the patient was holding).

One concern arises as to whether the striking tendency to generalize beyond the verb-specific input in both Perek & Goldberg (2015) and in the latter studies by Thothathiri and Rattinger (2016) was facilitated by prior knowledge of the sorts of information that individual verbs convey. The constructions used by Perek & Goldberg (2015) differed in terms of information structure, and adult participants can be expected to know that individual verbs are not generally associated with differences in information structure. In particular, whether a pronoun or a lexical noun phrase is appropriate in a given context is not something that usually depends on individual verbs. Relatedly, the two constructions used by Thothathiri and Rattinger (2016) differed in terms of what are normally considered adjuncts, and adjuncts are defined to be constituents that are not dependent on, or conditioned by, particular verbs. Therefore, in both cases, the remarkable tendency to generalize beyond verb-specific information in the input could have resulted from adults' understanding that the difference between the two constructions was not likely conditioned by individual verbs.

Experiment 1 is designed to address this concern. In both of the present studies, participants are exposed to two novel word order constructions that differ in terms of core clausal semantics. In particular, one construction exclusively describes actions that have a strong effect on the undergoer argument; the other construction describes actions with a weaker but otherwise similar effect. This contrast is just the sort of contrast that can readily be conveyed by distinct verbs (*tease* vs. *harass*; *charm* vs. *enchant*; *tap* vs. *smack*), and there is no English word order construction that designates this difference. Therefore, if participants extend (in a production task) and accept (in a judgment task) verbs for use in the alternative construction depending on whether the effect on the undergoer is strong or weak, it is not likely due to any prior knowledge that word order constructions should be more likely responsible for conveying the degree of affectedness than verbs.

Experiment 1 is similar in design to previous experiments by Wonnacott et al. (2008), Perek & Goldberg (2015) and Thothathiri & Rattinger (2016), in that it includes a lexicalist condition, in which each of six verbs is consistently witnessed in only one of the two constructions (three in each), and an alternating condition, in which two of the six verbs are witnessed in both constructions. When faced with a choice as to whether to ignore the perfectly predictable but functionless verb-specific distribution *or* the correlation between the semantics of the event type and construction, we hypothesize that participants will tend to ignore the verb-specific distribution in favor of generalizing verbs for use in whichever construction better conveys the intended message. The tendency to rely on the semantics of the event type to determine which construction is preferable is hypothesized to be especially strong in the alternating condition in which a minority of verbs are witnessed being used in both constructions.

2. Experiment 1

In a between-subjects design, participants were assigned to either the lexicalist condition or the alternating condition. In the lexicalist condition, each participant witnessed 3 verbs only occurring in the OSV construction accompanied by scenes in which unique actions were performed with weak effects on the undergoer; and 3 other verbs only occurring in the SOV construction accompanied by scenes in which unique actions were performed with strong effects on the undergoer. In the alternating condition, each participant witnessed 2 verbs only occurring in the OSV construction, 2 verbs only occurring in the SOV construction, and 2 verbs occurring in each construction 50% of the time. As in the lexicalist condition, all uses of the OSV construction in the alternating condition were matched with scenes displaying weak effects on the undergoer, and all uses of the SOV construction were matched with scenes displaying strong effects. At test, participants were asked to describe similar scenes that involved either a strong or a weak effect on an undergoer argument.

If speakers base their productions solely on the basis of distributional evidence in the input, we would expect speakers to restrict their productions to use each verb only in the construction in which it had been witnessed. If, however, speakers prefer to use constructions that are better suited to the discourse, speakers may display a tendency to disregard verb-specific distributional evidence in the input. It is also possible that speakers are capable of using both factors to some extent, as was the case in Perek & Goldberg (2015, Exp. 1)'s lexicalist condition. In this case, we might see a degree of lexical conservatism as well as some sensitivity to the functions of the constructions.

2.1 Participants

Participants were 24 undergraduate students at Princeton University. Eighteen of them participated in the experiment for course credit, and the other six received payment. Most participants were students at the Department of Psychology; some who received payment came from other departments. All were native speakers of English and had normal or corrected vision (16 female, 8 male, aged 18-22, mean 19.54).

2.2 Materials

Word order in the artificial language departed from standard English syntax, and consisted of two constructions involving different word orders: Subject Object Verb (SOV) or Object-Subject-Verb (OSV). A suffix *-po* was appended to the object noun in order to disambiguate between the two word orders (e.g., *the cat-po*). Each of six verbs appeared only in one construction or the other and each had a distinct meaning including: BLOW-ON (the agent bends over and blows air at the undergoer), HEADBUTT, KICK, PUNCH, PUSH, SLAP (with both hands), SPIN (the agent spins towards and hits the undergoer), SWIRL-STRIKE (the agent strikes the undergoer with a swirling blow).

The semantics of the two constructions were distinct in that the SOV construction always described actions that had strong effects on the undergoer (4) while the OSV construction always described actions that had weak effects (5):

(4) NP_{Agent} [the N-po]$_{Undergoer}$ V (SOV-strong effect on undergoer)
"agent acts on undergoer causing a strong effect"

(5) [the N-po]$_{Undergoer}$ NP_{Agent} V (OSV-weak effect on undergoer)
"agent acts on undergoer causing a weak effect"

Strong effects consisted in causing the undergoer to rapidly move all the way across the screen and out of the scene, while performing dramatic gestures like throwing their arms backwards, arching their back at a 90 degree angle, etc. Weak effects involved undergoers moving only slightly and performing similar but less ample gestures. At the end of strong-effect scenes, the undergoer was no longer visible on screen, while it remained visible in weak-effect scenes. This difference provided a visual cue for participants to distinguish the two different kinds of scene. Both versions of each action (strong effect and weak effect) were enacted by anthropomorphized animals in 3D animations recorded as video clips.[1]

The lexicon of the artificial language included six English names for animals (*cat, monkey, panda, pig, rabbit, wolf*) and eight nonce verbs: *glim, grash, moop, norp, pilk, speff, tonk,* and *wub*. Six of these verbs (randomly selected for each participant) were used in the exposure phase; the two remaining novel verbs were only used in the test phase, in order to assess how learners would treat items for which they did not receive any prior distributional information. The assignment of verb forms to the eight verb meanings described above was randomized for each participant.

As described in more detail below, participants were then asked to produce sentences to describe scenes that involved either strong or weak effects on the undergoer argument. We also collected acceptability ratings, as described subsequently.

2.3 Procedure

The experiment was programmed as a computer task implemented with PsychoPy (Pierce 2007) and run on a MacBook Pro laptop. All the instructions were given in written form on the computer screen. For each participant, the experiment was conducted over two sessions more than 24 hours and less than 48 hours apart.

Each session was divided into an exposure phase and a test phase. In the exposure phase, participants were gradually introduced to the artificial language. They were first shown a slowly rotating picture of each of the six animals involved in the stimuli scenes, paired with a description of the type "this is the panda/rabbit/etc." They were then exposed to the six verbs used in the exposure set by watching an example of each action (with randomly selected animal characters) paired with a description of the type "this is V-ing." Participants then proceeded to a vocabulary test that consisted in a forced-choice comprehension task: they had to identify each of the six verbs by choosing (by mouse click) which of two scenes designated a particular novel action named by one of the nonce verbs. Feedback (i.e., whether the answer was correct or not) was provided after each answer (thus allowing participants to refine their vocabulary knowledge). The vocabulary test ended when all six verbs were correctly identified twice in a row. This test was meant to ensure that all participants had a reasonable grasp of the verbal lexicon before exposing them to full sentences.

---

[1] The computer animations were created with Alice (http://www.alice.org), a visual programming language platform designed for educational purposes that allows users to create 3D-animated "virtual worlds" in which agents can be programmed to move and act in certain ways by means of a "point-and-click" interface.

In order to remain neutral as to how strong of an effect was involved for each verb, during the vocabulary learning phase, the effect was hidden from view during both the presentation of verbs and the vocabulary test: in the videos, a wall was seen sliding in front of the undergoer, right before the agent initiated the action. Importantly, the unique gestures performed by the agent for each verb were fully and clearly visible.

After completion of the vocabulary test, participants were exposed to sentences in the artificial language. They were shown three blocks of twelve scenes matched with a sentence description (thus totaling 36 input sentence-scene pairs), and were instructed to repeat each sentence out loud. Each of the six verbs was used twice in each block. The same pair of animals was used in all sentences of the exposure set, with balanced assignment to agent and patient role. This was done in order to focus participants' attention on the actions rather than on the arguments.

All sentence stimuli presented to participants in the exposure phase were displayed on the screen in written form and played in audio form on the laptop's speakers. The sentences were recorded into audio files by a computer-generated voice, by means of the MacinTalk text-to-speech synthesizer on Mac OS X 10.10, using the high-quality American English voice "Will" developed by Acapela Group and purchased through the Infovox iVox interface.

The test phase, described in detail below, included a production task in both sessions, followed by a sentence-rating task in session 2 only.

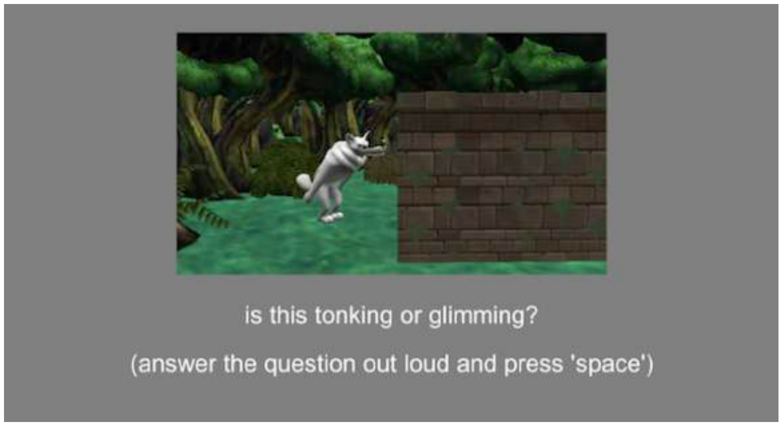
2.3.1 Production task

The production task contained 32 triples consisting of 1) a vocabulary question, 2) a sentence comprehension question and 3) a sentence production question (always in that order). The dependent measure of interest is the production data; the other tasks were meant to act as distractors and were intended to counter possible effects of self-priming.

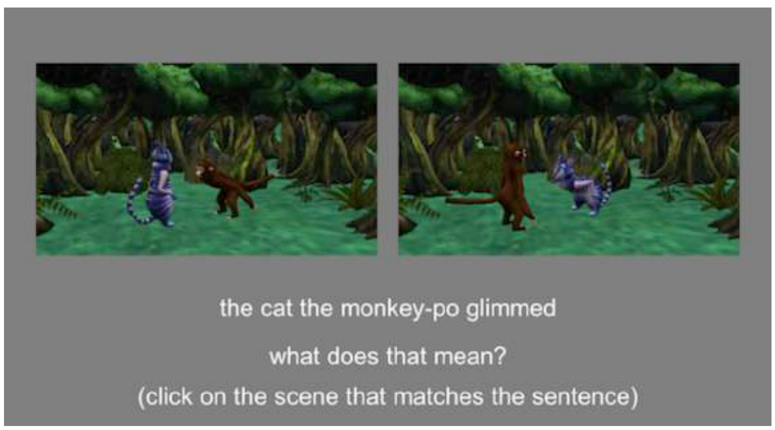
Sentence production task (question manipulation): Participants were prompted by the question *what happened here?* to describe a scene displayed on the screen by constructing a sentence in the artificial language. To facilitate the task, the verb was provided in written form (in the past tense) on the computer screen.

All six verbs introduced during the exposure phase, as well as two additional novel verbs, were presented four times during the production task, twice with a scene showing a weak effect on the undergoer, and twice with a scene showing a strong effect, each time with a different pair of agent and undergoer arguments. In all tasks, the left-to-right orientation of the undergoer and agent in the scene was randomly determined for each trial, with the agent presented on the right in half the scenes and on the left in the other half. The participants' responses to the production task in each trial were recorded using the laptop's microphone.

The two distractor tasks are described below, and an example triplet of tasks is illustrated by a screenshot in Figure 1.



Vocabulary distractor task



Comprehension distractor task

Production task

Figure 1: Screenshots of the three tasks given in each comprehension/production test triple. Testing consisted of 32 such triples.

Vocabulary distractor task: Participants were asked to identify the correct label for a given action shown on the screen (i.e., a verb) from two alternatives. For each trial, the two verbs were randomly selected from the six verbs used in the exposure phase, and the linear position of the right answer in the question was randomly determined. Participants had to provide their answers verbally but their responses were not recorded.

Sentence comprehension distractor task: In this task, participants were presented with a sentence and had to identify its meaning by choosing one of two scenes displayed on the screen. Each of the two constructions occurred equally often within the set of comprehension questions. The verb was randomly selected among those attested with the construction in the input, but it was always different from the one presented in the following production question. The two scenes displayed the same action and the same two characters, but they differed in terms of the assignment of thematic roles (the agent in the first scene was the undergoer in the second scene, and vice versa). The participants had to provide their answers by clicking on the matching scene with the computer mouse.

2.3.2 Sentence rating task

The sentence rating task was given to participants during session 2 only, following the production task. It consists in a standard acceptability judgment task. Participants were presented with 24 sentences paired with scenes and had to rate each sentence for acceptability given the target scene that it was supposed to describe. An example screenshot of the sentence rating task is showed in Figure 2.

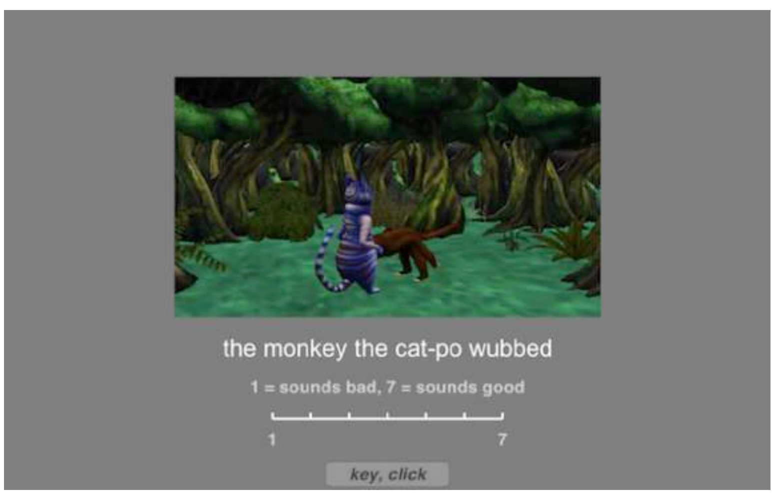


Figure 2: Example screenshot of the sentence rating task, with the verb *wub* used in the SOV construction.

Participants provided responses on a 7-point Likert scale, with 1 being "sounds bad" and 7 being "sounds good." All six verbs shown during the exposure phase were used four times each, once in each of the following combinations of sentence and type of scene: SOV with a strong effect on the undergoer and SOV with a weak effect; OSV with a strong effect on the undergoer, and with a weak effect. Participants were explicitly instructed to pay attention to not only whether the sentence made a well-formed string of words in the artificial language, but also importantly whether the meaning of the sentence matched the scene shown to them.

2.4 Results

Because we are interested in language use and not language learning per se, we focus below on the data collected after the second and final day of exposure, i.e., at the outcome of the learning process. We describe the results of the production and sentence rating tasks in turn.[2] Our entire dataset (including the data from both day 1 and day 2) is available as an online supplement.

2.4.1 Production task

The results of the production task were coded according to which construction was used. Sentences consisting of a regular noun phrase referring to the agent, a noun phrase followed by the particle *–po* referring to the undergoer, and the verb (in that order), were coded as "SOV". Sentences consisting of the same noun phrases in the opposite order (patient then agent) followed by the verb were coded as "OSV." 136 productions (amounting to 9% of the dataset) that did not fit either of these patterns were treated as errors and left out of the analysis. Notably, cases in which participants used the right order of arguments but attached the particle *–po* to the wrong noun phrase (i.e., the agent) were excluded. 18 responses (1.2%) failed to be recorded because the participant proceeded to the next trial before having fully uttered a sentence, or because of some other technical issue. For SOV and OSV sentences, misnaming one animal was ignored as long as the other animal was correctly labeled (thus allowing the thematic roles to be identifiable despite the error). In the event that the subject hesitated or produced multiple sentences, only their last full production was considered. Even though the correct verb was provided in each production trial, some participants occasionally uttered the wrong verb; these cases were also excluded. The coding procedure left us with 678 usable datapoints in the lexicalist condition, and 691 in the alternating condition.

The relative proportions of SOV and OSV productions are plotted in Figure 3, separately for each of the two conditions. The learned constructions were generally used appropriately in both conditions, albeit to varying extents: the OSV construction tended

[2]As intended, by day 2, performance on the comprehension task was at ceiling, in that participants were successfully able to assign thematic roles to the arguments of the verbs, identifying the correct scene 98.9% of the time.

to be used when the effect on the undergoer was weak, and the SVO construction tended to be used when the effect on the undergoer was strong. The same general trend is found for all verb types, regardless of how the verb was witnessed during exposure.

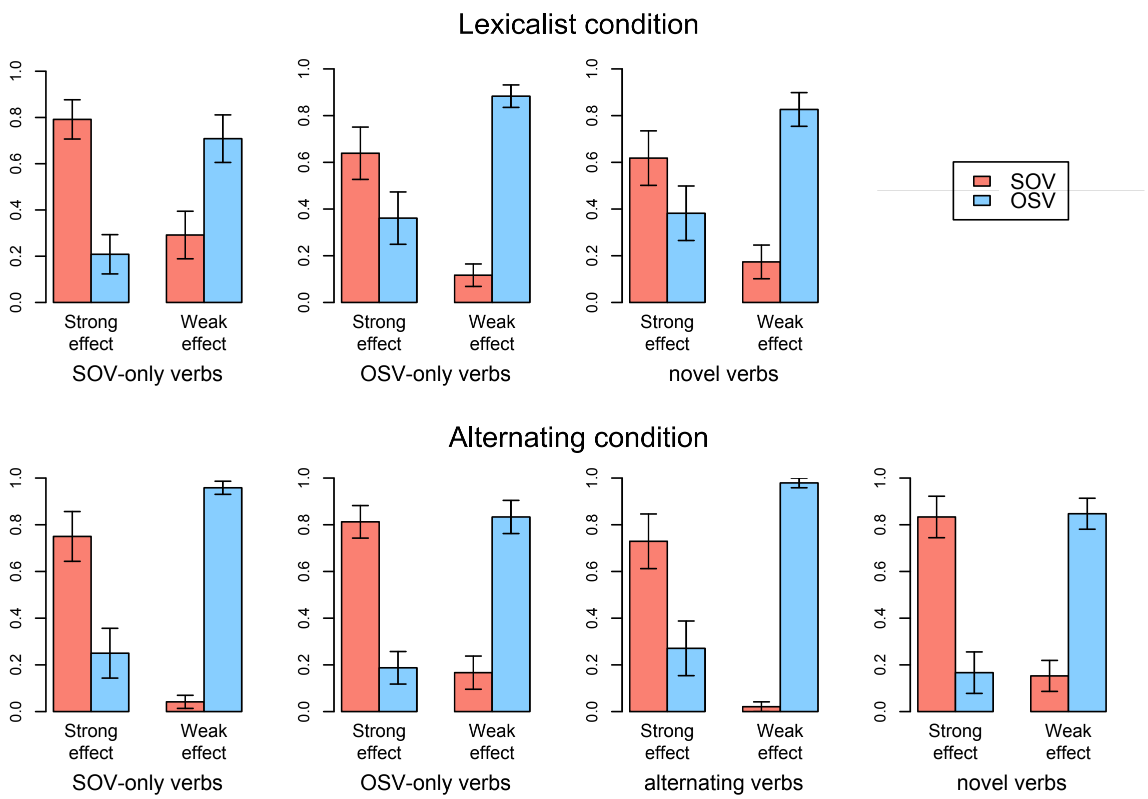


Figure 3: Results from Experiment 1. Each of the seven panels includes the proportion of participants' productions that describe scenes in which there was a strong effect on the undergoer (left side) or a weak effect on the undergoer (right side). Proportions of SOV and OSV productions for the verbs presented only in the SOV construction, only in the OSV construction, and for new novel verbs. The Alternating condition included two verbs that appeared in both constructions and performance on these is represented in the third panel on the bottom row.

To test for statistical significance, we submitted the data to mixed effects logistic regression, using the package lme4 in the R environment (Bates et al. 2011).[3] Each

[3] We used the 1.1-7 version of lme4. The *p*-values were calculated by the "summary" function from the package lmerTest version 2.0-25, which uses Satterthwaite's approximations to degrees of freedom (SAS Institute Inc., 1978). The R^2 values reported below each table were calculated with Nakagawa & Schielzeth's (2013) method, as

production of SOV or OSV is one observation in the dataset. The dependent variable, SOV (binary), records whether the utterance used the SOV construction vs. the OSV construction. In the regression model, we evaluate the factors that influence the production of one construction over the other, in particular as regards whether participants use these constructions productively, i.e., with verbs that were not witnessed in these constructions during exposure, or conservatively, i.e., with the same verbs that they were witnessed with in the input. For this reason, the data fitting the regression model does not include productions of sentences with alternating verbs, since it does not make sense to assess productivity with verbs for which the input provides explicit evidence that they can be used in both constructions. Also, keeping alternating verbs in the dataset would create empty cells and thus prevent the use of regression modeling if we are to include input condition as a factor, since such verbs are only found in the alternating condition.

There are three predictors (fixed effects) in the regression model:

a) Effect on the undergoer (Effect), a binary variable that captures whether the scene involved a strong or weak effect on the undergoer (strong vs. weak);
b) Verb type (VerbType), a categorical variable that captures whether a verb had been witnessed only in the SOV construction (SOV-only), only in the OSV construction (OSV-only), or not witnessed at all in the input (novel). The last case also serves as a baseline, since the input provided no reason to be biased toward one construction or another, and it is therefore used as the reference level for this factor in the regression model.
c) Input condition (Condition), a binary variable that indicates whether the participant was exposed to a lexicalist input, where each verb always occurs in the same construction, or to an alternating input, where two verbs are witnessed in both constructions (lexicalist vs. alternating).

Following standard model selection procedure, we first ran the most complex model containing all interactions between fixed effects, and proceeded stepwise by removing non-significant interactions one by one (Baayen 2008). The final model contains Condition, Effect, and VerbType as main effects, and the interaction between Condition and Effect. The fixed effects estimated by the fitted model are reported in Table 1.[4]

---

implement by the MuMIn package version 1.15.6.

[4] The full output of the lmer function for all mixed models discussed in this paper can be found in the online supplement.

Random effects for subjects (Subject), verb forms (Verb), and verb meanings (Meaning) were included in the model in order to factor in subject-specific preferences and to control for potential constructional biases that might happen to be inherently associated with particular verb forms or meanings.

We followed Barr et al.'s (2013) in starting with a maximal random effect structure containing random intercepts for Subject and by-participant random slopes for the factors Verb and Meaning. Novel verbs (i.e., not witnessed in the input) are taken as the baseline level for the factor VerbType. The model initially failed to converge, and only did so when we removed all random slopes, thus only keeping random intercepts for the three factors. It should be noted that the variance of Verb and Meaning is extremely small (below 0.0001),[5] which means that these factors had very little effect on the subjects' productions. The same random effect structure was used for all models reported in this paper, on the basis of the same criteria. We found a classification accuracy (i.e., the percentage of data points for which the model predicts the right construction) of 78.47%,[6] which indicates that the model is a reasonably good fit for the data.

	Estimate	Std. error	z-value	p-value
(Intercept)	-1.6520	0.2884	-5.728	< 0.0001 ***
Condition (lexicalist)	-0.1341	0.3196	-0.420	0.6748
Effect (strong)	3.4756	0.3509	9.906	< 0.0001 ***
VerbType (SOV-only)	0.8111	0.2529	3.208	0.0013 **
VerbType (OSV-only)	-0.0366	0.2510	-0.146	0.8842
Condition (lexicalist) x Effect (strong)	-1.1113	0.4226	-2.630	0.0085 **

Table 1: Fixed effects of the logistic regression model predicting the production of the SOV construction. Novel verbs (i.e., not witnessed in the input) are taken as the baseline level for the factor VerbType. Classification accuracy = 78.47%, Marginal R^2 = 40.93%, Conditional R^2 = 43.03%. Model formula: SOV ~ Condition * Effect + VerbType + (1 | Subject) + (1 | Verb) + (1 | Meaning).

Since uses of the SOV construction were coded as '1', positive values of the estimates in Table 1 indicate that the corresponding factor has a positive effect on the use of the SOV

[5] SD_{Subject} = 0.3475, SD_{Verb} = 0.0001, SD_{Meaning} = 0.0001.

[6] For this and all subsequent logistic regression models, the classification accuracy is reported in the legend of the relevant table.

construction, and conversely, negative values indicate that the factor favors the use of the OSV construction.

As was evident in Figure 3, which construction a verb was used with depended strongly on the semantics of the scenes involved: i.e., whether the undergoer was strongly or weakly affected. Accordingly, we find a strong and significant main effect of the factor Effect in the regression model; this effect is positive for the level 'strong', confirming that the SOV construction is significantly more likely to be used in the presence of a strong effect on the undergoer. In fact, Effect had a significant impact on how a verb was used, regardless of how that verb had been witnessed in the input; that is, interactions of Effect with the levels of VerbType were not significant when included in the model (Condition(lexicalist) x VerbType(SOV-only): $\beta = 0.2916$, $SE = 0.5053$, $z = 0.577$, $p = 0.5638$; Condition(lexicalist) x VerbType(OSV-only): $\beta = -0.2264$, $SE = 0.5041$, $z = 0.449$, $p = 0.6533$).[7]

At the same time, the significant interaction of Condition and Effect indicates that the impact of Effect was somewhat less pronounced in the lexicalist condition than in the alternating condition (since the estimate of the interaction effect is negative). There is also a significant difference between novel verbs and SOV-only verbs, as shown by the positive main effect of VerbType (SOV-only), with the latter favoring SOV productions. OSV-only verbs, however, do not significantly differ from novel verbs. Changing the reference level of VerbType to SOV-only also reveals a significant difference between SOV-only verbs and OSV-only verbs ($\beta = -0.8477$, $SE = 0.2365$, $z = -3.584$, $p = 0.0003$).

In sum, participants were more likely to use SOV-only verbs in the SOV construction at test. But this effect of lexical conservatism is rather weak compared to that of constructional meaning, and is only evident for verbs witnessed in the SOV construction during exposure. Moreover, the effect does not significantly vary according the input condition, as the interactions between Condition and the levels of VerbType were not significant (and very weak) when added to the model, and the corresponding effects (Condition (lexicalist) × VerbType (SOV-only): $\beta = 0.0527$, $SE = 0.5218$, $z = 0.103$, p = 0.9196; Condition (lexicalist) × VerbType (OSV-only): $\beta = -0.0752$, $SE = 0.5182$, $z = -0.145$, $p = 0.8846$).

To summarize, participants in both input conditions were sensitive to the functions of the newly learned constructions: they readily extended verbs for use in either construction, depending on whether the effect on the undergoer was strong or weak. Whether verbs had only been witnessed in the SOV-strong or the OSV-weak

[7] The model failed to converge when the interaction term was added, unless we removed the random intercept for Verb. The figures reported here are from the latter model. The same observation applies for the model including the interaction between Condition and VerbType (see below).

constructions during exposure had little impact on their choice of construction in either condition.

2.4.2 Sentence rating task

Data from the sentence rating task is consistent with the production results. In accord with standard practice in grammaticality rating studies, we converted raw ratings on the 7-point scale to z-scores in order to control for the fact that subjects often use the scale in different ways. The conversion to z-scores replaces each rating by a value that indicates how many standard deviations it diverges from the subject's average rating. Z-scores are calculated by subtracting the mean of all ratings provided by the same participant from each of that participant's original 7-point-scale ratings, and dividing the difference by the standard deviation of these ratings.[8]

Figure 4 presents the distributions of z-scores in the two input conditions in the form of box plots. The distributions are grouped by how the verb was witnessed during exposure: SOV-only, OSV-only, and alternating verbs (in the alternating condition only). Each boxplot is further divided into the four possible combinations of construction and effect on the undergoer found in the stimuli set, from left to right: SOV with strong effect, OSV with weak effect, SOV with weak effect, and OSV with strong effect. The first two are combinations attested in the input: we call them "effect-congruent", in the sense that the scene described by the sentence matches observed usage of the construction as far as the effect on the undergoer is concerned. The latter two conflict with the input in the same respect, and are therefore called "effect-incongruent". For better visualization, the distributions of effect-congruent combinations are colored in green, and those of effect-incongruent combinations in red.

As is standard in box-plots, the boxes are delimited by the lower and upper quartiles of each distribution; in other words, they correspond to the middle range and contain half the values of the distribution. The black stripe is the median: each half of the distribution is located to the top and bottom of this value, which can thus be taken as an indication of the central tendency. The dashed lines ending with whiskers represent values that are outside the lower and upper quartiles but still within 1.5 times the

[8] One subject had to be excluded from the analysis because they provided the same rating for all sentences (7, i.e., full grammaticality). Consequently, their z-scores could not be calculated, because the standard deviation of their ratings, used as divisor in the calculation, equals to 0, since there is no variation. The final dataset analyzed in this section totals 552 observations.

interquartile range (i.e., the difference between the upper and lower quartiles). The values outside this range are outliers and represented by bullets in the plots.

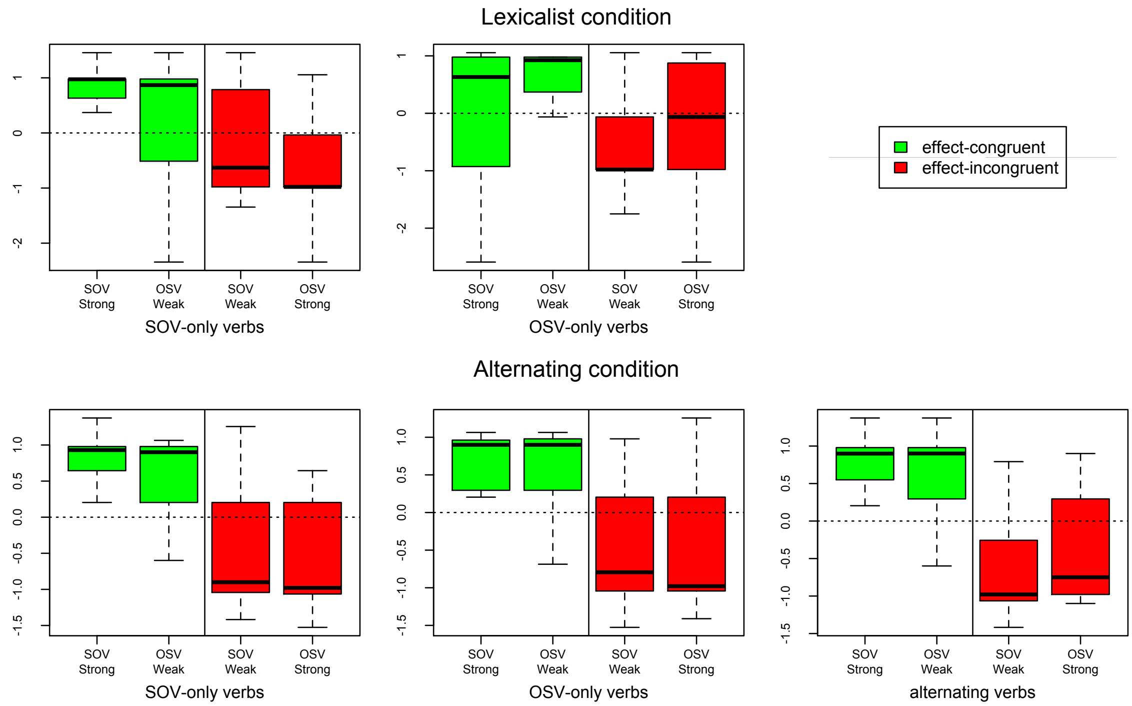


Figure 4: Box plots of the distribution of grammaticality ratings (z-scores) provided by participants in the lexicalist condition (top) and in the condition in which one third of verbs alternated (bottom), for each verb type, and each combination of construction and effect on the undergoer seen in the stimuli. Combinations that were congruent with the input as regards the effect on the undergoer (i.e., SOV with strong effect, OSV with weak effect) are plotted on the left-hand side of each box and colored in green; the other, incongruent combinations are plotted on the right-hand side of each box and colored in red. Outliers are not plotted.

As can be seen in Figure 4, participants in both the lexicalist and the alternating condition generally judged as more acceptable constructions that described congruent scenes: the SOV construction when the effect on the undergoer was strong, or the OSV construction when the effect was weak. The lexicalist condition shows a small effect of how the verb involved was witnessed in the input, with a broader range of scores evident when the verb is used in the construction that had not been witnessed during exposure. In the alternating condition, there is no effect of how the verbs had been witnessed occurring in the input: instead, participants fully generalize each of the two constructions for use in appropriate scenes with any verb.

To test whether these differences are significant, we submitted the sentence ratings to mixed effects linear regression. The regression model contains three predictors: (i) EffectCongruent, a binary variable recording whether the construction is used to describe a scene with the same kind of effect as in the input, (ii) VerbCongruent, a binary variable which records whether the verb is used in a construction with which it was witnessed in the input, and (iii) Condition, a binary variable that records which input condition the participant was exposed to (as in Section 2.3.1). EffectCongruent is set as true if the construction is SOV and the effect is strong, or if the construction is OSV and the effect is weak, and false otherwise. VerbCongruent is set as true for SOV-only verbs used in the SOV construction, for OSV-only verbs used in the OSV construction, and for alternating verbs used in either construction (in the alternating condition only), and it is set as false for SOV-only verbs used in the OSV construction and OSV-verbs used in the SOV construction. As previously, by-subject, by-verb-form, and by-meaning random effects were also included in the regression, but none of them captured significant variance (all < 0.0001).

The fixed effects of the regression model are reported in Table 2. We find a positive main effect of EffectCongruent, but no main effect of VerbCongruent. However, both predictors are involved in significant interactions with Condition. In the lexicalist condition, EffectCongruent had a negative impact on grammaticality judgments; in other words, participants in that condition had a lower tendency to assume that the effect on the undergoer was a critical factor when rating instances of each construction than participants in the alternating condition, in line with what was found in the production task. Contrary to the production task, however, an effect of lexical conservativeness is detected in the positive interaction of VerbCongruent with Condition, showing that participants in the lexicalist condition tend to regard sentences using a verb in the same construction it was witnessed with in the input slightly more favorably than participants in the alternating condition. Yet, this effect is quite weak compared to the main effect of EffectCongruent, with which it can hardly compete; moreover, the fact that it reaches significance is probably due to the coding used for the sentence rating dataset, which gives more statistical power. Hence, the pattern of results is by and large similar to that of the production task.

	Estimate	Std. error	t-value	p-value
(Intercept)	-0.6170	0.1031	-5.988	< 0.0001 ***
EffectCongruent (true)	1.1210	0.1031	10.878	< 0.0001 ***
VerbCongruent (true)	0.0848	0.1093	0.775	0.4384
Condition (lexicalist)	-0.0147	0.1339	-0.110	0.9127
EffectCongruent (true) x Condition (lexicalist)	-0.2936	0.1427	-2.058	0.0401 *
VerbCongruent (true) x Condition (lexicalist)	0.35123	0.1472	2.385	0.0174 *

Table 2: Fixed effects of the linear regression model predicting the z-score ratings provided by subjects in the sentence rating task in Experiment 1. Marginal R^2 = 27.48%, Conditional R^2 = 27.48%. Model formula: Zscore ~ EffectCongruent * Condition + VerbCongruent * Condition + (1 | Subject) + (1 | Verb) + (1 | Meaning).

To summarize, the results of the sentence rating task are consistent with those of the production task. In the lexicalist condition, sentences were judged to be somewhat less grammatical when they contain a verb used in a different construction from the one it had been attested with in the input, although the median score is just as high in either case. In the condition in which one third of the verbs are witnessed alternating, all verbs are judged equally grammatical in either learned construction, as long as the effect on the undergoer was appropriately either strong or weak

2.5 Discussion

The main aim of Experiment 1 was to address the concern that the strong tendency to generalize on the basis of the functions of the constructions that had been found in Perek & Goldberg (2015, Exp.1) and Thothathiri & Rattinger (2016, Exps 2 & 3) was due to the fact that participants in these earlier studies may have relied on prior knowledge that neither information structure nor adjuncts are generally associated with individual verbs. The present experiment instead varied whether an effect on the undergoer was strong or weak, as this type of change is readily associated with distinct verbs (e.g., *crush vs. pulverize; edit vs. rewrite; wipe vs. scrub*). Therefore, prior knowledge was not expected to lead learners to generalize a familiar verb for use in a distinct construction in order to convey a stronger or weaker effect on the undergoer argument. Nonetheless, just as in the earlier studies, participants did generalize beyond the lexically specific input, even in the lexicalist condition in which they did not witness any of the verbs alternating. In fact, participants showed little evidence of lexically conservative behavior in the lexicalist condition, and *no* evidence of lexically conservative behavior in the alternating condition in which only two of the six verbs witnessed occurred in both constructions.

A second contribution of Experiment 1 is that it allows us to contrast two different interpretations of *why* participants are likely to generalize beyond their input when two constructions are assigned distinct functions. While Perek & Goldberg (2015) emphasized the idea that learners prefer to select the construction which better suits the discourse context and therefore affords more expressive power, Thothathiri and Rattinger (2015) interpreted their parallel findings in terms of the relative *cue reliability* of verbs *vs.* types of scenes in predicting which construction was used during exposure. In particular, in Thothathiri and Rattinger's first experiment, the two novel transitive constructions were identical in function and each was used equally often; therefore the probability that a given transitive scene would be described by either construction was .50. At the same time, two out of 12 verbs were witnessed in both constructions, while 10 verbs uniquely (and arbitrarily) predicted which construction was witnessed during exposure. Therefore, the cue reliability of verbs was 1 for 5/6 of the verbs and was .50 for 1/6 of the verbs for an average cue validity across verbs of .92 for predicting the choice of construction. In this first experiment, Thothathiri and Rattinger found that participants were lexically conservative, and argued that this was a result of the fact that the verbs better predicted which construction was used than the type of scene had during exposure.

In two additional experiments, like Perek & Goldberg (2015, Exp.1), Thothathiri and Rattinger (2016) assigned distinct functions to the two constructions. Ten verbs witnessed during exposure either occurred exclusively in one construction or the other (5 verbs in each) and two verbs alternated between the two constructions. Thothathiri and Rattinger noted that the average cue reliability of verbs remained the same as in their first experiment, but the two different constructions were uniquely associated with distinct types of events (a scene involving an instrument *or* a scene with a modifier). They found that all verbs were generalized for use in the construction that was appropriate to express the intended message, replicating the generalization effect found by Perek & Goldberg (2015, exp. 1). They suggested that participants generalized beyond the verb-specific input *because* the type of scene (whether an instrument or modifier was relevant) predicted which construction was used better than the verbs did.

To summarize, while the transitive events were less predictive of which construction would be used than the verbs in Thothathiri and Rattinger (2016)'s first experiment, they also did not serve distinct communicative functions. The question therefore is, were participants more lexically conservative in Thothathiri and Rattinger (2016) first experiment relative to the second one—and in Perek & Goldberg (2015)'s second experiment relative to their first one—because the verbs were more predictive of which construction would be used, or because there was no communicative motivation to generalize the verbs for use in the alternative construction when the two constructions served the identical function?

In the present lexicalist condition, the cue validities that Thothathiri & Rattinger (2016) had appealed to are matched. In particular, during exposure, verbs predicted which construction was used with a probability of 1, *and* the degree of effect on the undergoer also predicted which construction was used with a probability of 1. But instead of relying equally on verbs and the type of scenes in choosing which construction to use, as would be predicted by an account based wholly on cue-validity, participants demonstrated a much stronger tendency to allow the type of scene to determine the choice of construction than to use the functionless verb-specific distribution. This was also true in the lexicalist condition of Perek & Goldberg (2015, Experiment 1). This suggests that when cue-validity is controlled for, the ability to convey an additional aspect of meaning, offered by the choice of construction, trumps the desire to simply obey the formal properties of the input.

At the same time, a comparison of the lexicalist and alternating conditions in the present experiment provides some evidence in favor of cue reliability as an additional factor, as Thothathiri & Rattinger 2016 had proposed. In particular, participants were even more likely to use the type of scene to predict which construction to use in the alternating condition, where the semantics of the scenes, but not the verbs, were perfectly predictive of which construction would appear during exposure. (Recall, in the alternating condition, two of the six verbs were witnessed in both constructions making verbs less than perfectly reliable cues). This was again, also true in a comparison of Perek & Goldberg (2015)'s lexicalist and alternating conditions. Interesting, as emphasized by Wonnacott et al. (2008), the relevant cue reliability is not determined by individual verbs, but by the statistics of the language overall: four of the verbs *were* perfect predictors of which construction would be used during exposure, but since two of the verbs alternated, participants completely ignored the verbs' distribution and freely used whichever construction better matched the scene at test. To summarize, the cue validity of verbs and scenes in the language at large seems to matter in which is relied on to select an appropriate construction, and when there is no difference in cue validity, speakers rely more on the scenes, overriding verb-specific distributions to gain more expressive power.

A concern might be raised that participants in Experiment 1 simply failed to learn which construction was associated with each verb in the input. Perek & Goldberg (2015; Experiment 2) had demonstrated that learners are capable of learning the verb-specific distribution of six transitive verbs with the present amount of exposure, which mitigates against this concern, but it is possible that the semantics of the constructions somehow interfered with learning the distribution of individual verbs. We will see that this concern is addressed by the results of Experiment 2.

The demonstration of widespread generalization on the basis of the functions of the constructions raises the issue of how generalizations are constrained, since it is clear that speakers do not extend real verbs for use in various constructions willy-nilly. That is, even productive alternations often have lexical exceptions (Braine 1970; Baker 1979; Pinker 1989; Levin 1993). For example, it is well-known that certain verbs in English resist occurring in the double-object construction, even though they would be perfectly interpretable and even when the construction's information structure properties would seem to be appropriate. That is, native English speakers disprefer the sentences in (6) in favor of a different construction, the *to*-dative or "caused-motion" construction in (7) (e.g., Pinker 1989; Levin 1993; Goldberg 1995; see Ambridge et al. 2014 for judgment data confirming the dispreference of examples such as those in (6) *vis a vis* those in (7):

(6) a. ?? She explained me something.
    b. ?? She dragged him the piano.
    c. ?? She mumbled him something.

(7) a. She explained something to me.
    b. She dragged the piano to him.
    c. She mumbled something to him.

It has been proposed that speakers essentially learn to avoid the type of formulations in (6) in favor of the formulations in (7) in the same way that speakers learn irregular morphological forms, e.g., *feet* is used instead of *foots*. We know that the latter are learned because learners systematically witness *feet* in contexts that would otherwise be appropriate for *foots* (Aronoff 1976; Kiparsky 1982). In the same way, if learners consistently witness the verbs *explain, drag* and *mumble* in the *to*-dative in contexts that would otherwise seem to favor the double-object construction, *to*-dative uses of these verbs may come to *statistically preempt* double-object uses of those verbs (Goldberg 1995). Previous work that had found evidence in favor of statistical preemption had used production or judgment data of familiar English constructions (Brooks & Tomasello 1999; Boyd & Goldberg 2011; Robenalt & Goldberg 2015). For example, Brooks & Tomasello (1999) found that novel verbs witnessed intransitively were preempted from being used transitively if a periphrastic causative is witnessed. Boyd & Goldberg (2011) found that novel "a-adjectives" that were witnessed in a relative clause use avoided being used prenominally, just as familiar a-adjectives are (e.g., *??the afraid boy*) (but see Yang 2015;and response by Goldberg & Boyd 2015).

The use of familiar constructions raises the possibility that learners brought with them prior knowledge that these particular constructions had item-specific properties. While this item-specificity itself may have been learned via statistical preemption, as the earlier studies assumed, we cannot rule out the possibility that the item-specific nature of those constructions was recognized by some other means. In order to address this issue, in Experiment 2, we use the novel constructions from Experiment 1—which we have already seen can be readily generalized—and investigate whether statistical preemption is used by speakers to learn the item-specific behavior of one verb, and whether the item-specific behavior of this one verb is generalized to other verbs that are learned concurrently.

3. Experiment 2

Experiment 2 again introduces six new verbs, each restricted to one of the two constructions used in Experiment 1, but in this case, one verb is witnessed consistently in the OSV construction *in contexts that vary as to whether the effect is strong or weak*. The other verbs are, as in Experiment 1, witnessed only in contexts congruent with the semantics of the construction they were assigned to. We hypothesize that the use of a verb in one construction in both semantic contexts will statistically preempt the use of that verb in the SOV construction, and therefore serve to constrain the constructional generalization. If so, this will provide support for idea that statistical preemption allows learners to avoid overgeneralizations without prior knowledge of a restriction. Since we know from previous results that learners tend to use the statistics of the language as a whole, we hypothesize that any restriction learned for the single verb that is witnessed in both contexts may be generalized to apply to other verbs to some extent as well. That is, we hypothesize an increase in lexically-specific behavior when results are compared with Experiment 1.

2.6 Participants

Participants in Experiment 2 were 12 undergraduate students at Princeton University (6 female, 6 male, aged 18-27, mean 20.17). All of them received course credit for their participation.

2.7 Materials

The exposure set contained the same number of sentences and nonce verbs as in Experiment 1. The assignment of verbs to constructions was identical to that of the lexicalist condition of Experiment 1: three verbs occurred exclusively in SOV, which is the construction associated with a strong effect on the undergoer argument (SOV verbs), and the other three verbs occurred exclusively in OSV, the construction associated with a weak effect (OSV verbs). The tasks were also identical to those used in Experiment 1, as were almost all other details regarding the artificial language and the exposure set.

The one key difference is that in this second experiment, a single OSV verb was witnessed in both types of semantic contexts: contexts yielding a strong effect on the undergoer and contexts yielding a weak effect on the undergoer. In this way, the weak-effect semantics associated with the OSV construction was probabilistic, holding for all scenes associated with two verbs and half of the scenes associated with a third verb. In order to ensure that the distinctive behavior of the preempted OSV verb was detectable by participants in the exposure phase, two instances of the preempted verb were presented at the very beginning of the set, and another four sentences at the very end of the exposure set. The verb was witnessing with a weak and a strong effect in succession, always in the same OSV construction. The semantics associated with the SOV construction was uniform: it was always associated with scenes that involved strong effects on the undergoer.

2.8 Procedure

The procedure was identical to the one used in Experiment 1.

2.9 Results

Production task

The same coding scheme was used as in Experiment 1. There were a total of 372 usable data points, after eliminating two productions that did not qualify as valid instances of either construction according to the coding criteria,[9] and ten further productions (2.6%) which failed to be recorded.

The proportions of SOV and OSV constructions in the subjects' productions are represented in Figure 5, with the 3 SOV-only verbs, all witnessed having a strong effect on the undergoer argument; 2 OSV-only verbs witnessed having a weak effect on the undergoer; 1 (OSV-only) verb witnessed having both a strong and a weak effect on the undergoer; and 2 new novel verbs.

[9]Performance in the comprehension task was comparable to that of Experiment 1. Participants identified the correct scene 96.4% of the time on average.

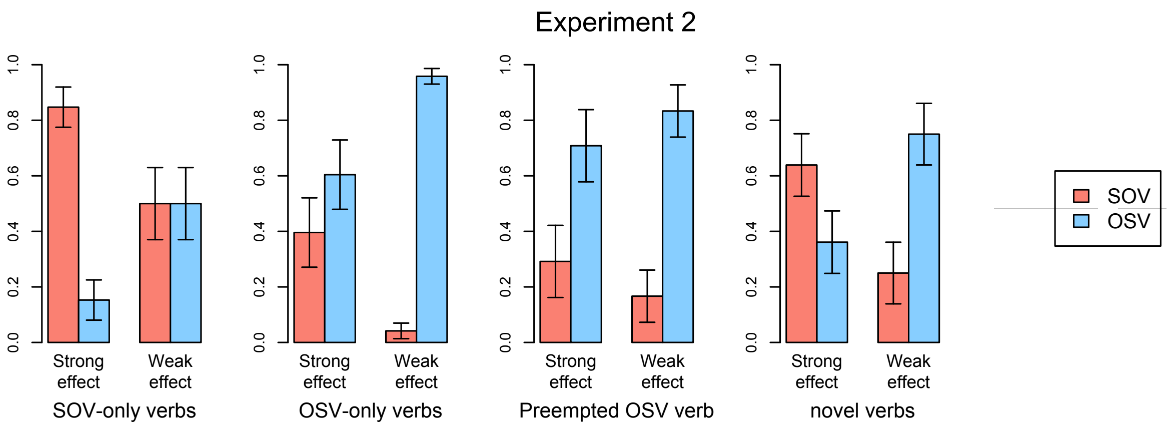


Figure 5: Proportions of SOV and OSV productions in Experiment 2

If we consider the new novel verbs first (far right panel of Figure 5), it is evident that the functions of the two constructions were readily detected, as they were in Experiment 1, since these verbs were used in whichever construction was better suited to the semantics of the scene being described, even though the function associated with the OSV construction was only probabilistic.

Of special interest is that, unlike any verbs in Experiment 1, the preempted OSV verb, which had been witnessed in both semantic contexts, shows a clear tendency to only be used in the OSV construction, regardless of context (3rd panel). Thus, witnessing this verb in the OSV construction, even when the undergoer was strongly affected, appears to have statistically preempted this verb's use in the SOV construction. Moreover, participants tended to treat *all* OSV verbs more lexically conservatively. That is, participants showed a tendency to use verbs that had only been witnessed in the OSV construction with congruent (weak effect) semantics to be used only in the OSV construction, even when the undergoer was strongly affected. Verbs which had only been witnessed in the SOV construction also displayed a tendency towards lexical conservativeness, in that they tended to use the SOV construction when the undergoer was strongly affected, but they generalized to the unwitnessed OSV construction half the time when the undergoer was weakly affected.

The data collected for Experiment 2 was fitted to a mixed effects logistic regression model similar to the one used in Experiment 1, which predicts the occurrence of the SOV construction from the fixed predictors Effect and VerbType, with random intercepts for Subject, Verb, and Meaning.[10] As in Experiment 1, novel verbs were treated as the baseline and used as the reference level of VerbType. We performed the

[10] The standard deviations of the random effects were as follows: SD_{Subject}=0.9292, SD_{Verb}=0.0015, SD_{Meaning}=0.1198.

same model selection procedure as in Experiment 1. Since no significant interactions between Effect and any of the levels of VerbType were found, we removed the interaction term from the model definition. The fixed effects of the final model are reported in Table 3. Note that the model with interactions failed to converge unless all random effects but Subject were removed, but even then, none of the interactions of VerbType with Effect were significant.

	Estimate	Std. error	z-value	p-value
(Intercept)	-1.3725	0.4172	-3.289	0.0010 **
Effect (strong)	2.0433	0.2928	6.978	< 0.0001 ***
VerbType (SOV-only)	1.3727	0.3686	3.724	0.0002 ***
VerbType (OSV-only)	-1.2858	0.4011	-3.205	0.0013 **
VerbType (preempted OSV)	-1.4558	0.4828	-3.015	0.0026 **

Table 3: Fixed effects of the logistic regression model predicting the occurrence of the SOV construction in the preemption condition. Classification accuracy = 76.34%, Marginal R^2 = 36.86%, Conditional R^2 = 50.05%. Model formula: SOV ~ Effect + VerbType + (1 | Subject) + (1 | Verb) + (1 | Meaning)

We find a significant positive main effect of Effect, showing that, when all verb types are collapsed, participants tended to use the SOV construction when the undergoer was strongly affected. However, this tendency is counterbalanced by significant effects of lexical conservativeness for each verb type: SOV-only verbs tended to be used in the SOV construction (as shown by the positive estimate), and the OSV verbs (OSV-only verbs and the preempted verb) tended to be used in the OSV construction (as shown by the negative estimate). When the reference level for VerbType was changed to "preempted OSV", the model showed no significant difference with OSV-only verbs (β = 0.1701, SE = 0.5090, F = 0.334, p = 0.7382), showing that the degree of lexical conservativeness was not measurably different between the two types of OSV verbs.

It is clear that witnessing a single verb used in both constructions, strongly mitigated the tendency to generalize on the basis of the constructions' semantics that we had seen in Experiment 1. This is consistent with Wonnacott et al.'s (2008) observation that learners decide how to use particular items partly on the basis of how other items have been witnessed being used. Here we find evidence that all verbs are somewhat more lexically conservative when there is evidence that a single verb loyally appears in the OSV-construction irrespective of whether the effect on the undergoer was strong or weak. The results of Experiment 2 are consistent with an interpretation in terms of cue-reliability, since participants tended to be lexically conservative when verbs predicted the choice of construction perfectly (probability = 1) and the effect on the undergoer only predicted the choice of construction with a probability of .92.

At the same time, participants were not entirely lexically conservative. That is, while they showed a tendency to respect the verb-specific input, they also displayed some tendency to generalize. In particular, participants were more likely to use the SOV construction—always associated with a strong effect—when the context portrayed a strong effect, than they were when the context portrayed a weak effect, and they were more likely to use the OSV construction—probabilistically associated with a weak effect—when the context they were describing involved a weak effect, than they were when the context involved a strong effect. This tendency was particularly evident in participants' productions with novel verbs, where participants strongly tended to produce the SOV construction when there was a strong effect on the undergoer and the OSV construction when the effect on the undergoer was weak.

2.8.1 Sentence rating task

All 288 data points collected in the sentence rating task in Experiment 2 were used in the analysis. The results are presented in Figure 6 in the form of box plots of the z-scores for each combination of construction and effect, plotted separately for each of the three verb types. As previously, congruent combinations are placed to the left of each plot and colored in green, and incongruent combinations are placed to the right and colored in red.

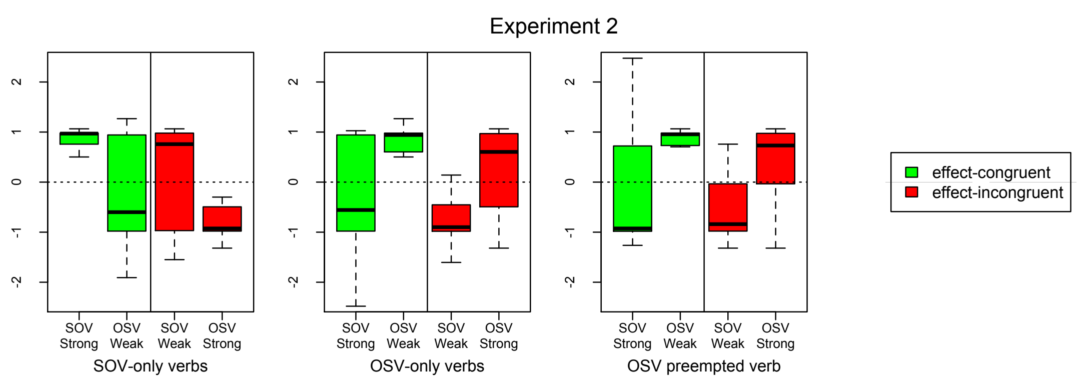


Figure 6: Experiment 2. Box plots of the distribution of grammaticality ratings (z-scores) provided by participants for each verb type, and each combination of construction and effect on the undergoer seen in the stimuli. Combinations that were congruent with the input as regards the effect on the undergoer (i.e., SOV with strong effect, OSV with weak effect) are plotted on the left-hand side of each box and colored in green; the other, incongruent combinations are plotted on the right-hand side of each box and colored in red. Outliers are not plotted.

All three types of verbs were judged to be more acceptable when they are used the way they had been witnessed, regardless of the effect on the undergoer; i.e., SOV-only verbs tend to be judged acceptable in the SOV construction and unacceptable in the OSV construction, and vice versa for all OSV verbs. An effect of congruency is only evident in the range of scores, as ratings for the verb in an unwitnessed construction spread higher when the construction was used with a congruent scene (strong effect for the OSV construction and weak effect for the SOV construction), and lower when the construction was used with an incongruent scene.

To test for significance, we submitted the sentence rating data of Experiment 2 to mixed effects linear regression, as in Experiment 1. The dependent variable in the model is the z-score sentence rating submitted for each trial, and the main predictors are EffectCongruent and VerbCongruent, two binary variables that respectively indicate whether the trial sentence uses the construction that is congruent with the effect on the undergoer as displayed in the accompanying video (SOV for strong effect and OSV for weak effect), and whether the verb is used in the construction it was consistently witnessed with in the input (SOV for SOV-only verbs and OSV for both OSV-only verbs and the preempted OSV verb). Subject, Verb, and Meaning were again included as random factors, but they did not capture significant variance (all < 0.0001). No significant interaction was found between EffectCongruent and VerbCongruent (β = -0.0134, *SE* = 0.2043, *t* = -0.066, *p* = 0.9476); hence it was removed from the final model. The fixed effects of the regression model are reported in Table 4.

	Estimate	Std. error	t-value	p-value
(Intercept)	-0.6172	0.0883	-6.987	< 0.0001 ***
EffectCongruent (true)	0.3859	0.1020	3.784	0.0002 ***
VerbCongruent (true)	0.8484	0.1020	8.319	< 0.0001 ***

Table 4: Fixed effects of the linear regression model predicting the z-score ratings provided by subjects in the sentence rating task in Experiment 2. Marginal R^2 = 22.54%, Conditional R^2 = 22.54%. Model formula: Zscore ~ EffectCongruent + VerbCongruent + (1 | Subject) + (1 | Verb) + (1 | Meaning)

Both types of congruency have a significant positive effect on sentence ratings. However, the effect of VerbCongruent is markedly stronger than that of EffectCongruent, as shown by the higher estimate. In other words, in Experiment 2, participants relied substantially more on how the verb had been witnessed during exposure than on the semantics associated with the constructions. We tested for interactions between the two kinds of congruency and VerbType, but none of them turned out significant, showing that there are no measurable differences across verb types in the strength of the effect of either kind of congruency.

In sum, the results of the sentence rating task are in line with those of the production task. While participants showed some degree of reliance on the match between construction and context when judging the acceptability of sentences, they largely tended to be lexically conservative with all verbs when the input contained a single verb consistently used in the OSV-construction regardless of whether the effect on the undergoer was strong or weak.

3. General discussion

The results of Experiment 1 confirm the idea that participants readily learn the functions of individual constructions, and readily extend verbs for new uses in an unwitnessed construction if the function of that construction is better suited to convey the intended message. When both verbs and the type of event (strong or weak effect on the undergoer) perfectly predict which construction is used during exposure—the lexicalist condition—participants displayed a strong tendency to use the construction that better suited the type of scene. This behavior was communicatively useful because the semantic contribution of verbs and constructions was independent of one another and additive. That is, each verb conveyed a specific kind of action and each construction conveyed whether the effect on the undergoer was strong or weak. If participants had instead obeyed the distributional properties of each verb, they would have been unable to convey the systematic differences in the degree of affectedness of the undergoer. In the alternating condition of Experiment 1, two verbs were unreliable predictors of which construction would be used, as both were used equally often in the two constructions. In this case, participants *entirely* ignored the distribution of all six verbs in the input, and used them all freely in either construction, dependent only on whether the effect was strong or weak.

Results from Experiment 2 added important nuance to the finding that participants readily generalize beyond their input. When learners witnessed a verb being used in the OSV construction to describe *either type of message*, they restricted that verb to the OSV construction for either type of message. The verb was thus statistically preempted from being used in the alternative construction. Participants in fact generalized this behavior to all verbs, preferring them in their respective constructions in both production and judgment tasks. The increase in verb-specific behavior, however, did not prevent participants from recognizing the semantics associated with each construction, as evidenced by their productions involving new novel verbs.

Importantly for the interpretation of Experiment 1, the results of Experiment 2 also demonstrate that participants were *capable* of learning the verb-specific biases that existed in the present experiment, even though they generalized beyond the verb-specific exposure in Experiment 1. Therefore, the results of Experiment 1 stand as an indication that speakers are willing to extend verbs for use in different constructions when doing so provides them with additional expressive power. This finding goes beyond previous related work in that the productive use of the constructions is not likely due to any prior assumption that constructions should be more likely to encode degree of affectedness than verbs.

The results of Experiment 2 lend important new support for the idea that learners are sensitive to the contexts in which particular verbs and constructions are used: witnessing one verb in the same construction regardless of whether the scene involved a strong or weak effect led learners to strongly tend to use all verbs in whichever construction they had been witnessed, even though participants demonstrated an appreciation of the functions of the constructions, particularly with new novel verbs. The tendency to generalize from evidence that a single verb is statistically preempted from occurring in a different construction to other verbs is striking evidence of the power of statistical preemption, especially given the fact that learners had no prior knowledge of whether to expect that verbs would be lexically restricted.

Additional work is needed to investigate whether children are as sensitive to the functions associated with abstract constructions as adults have been found to be, and whether they are as sensitive to statistical preemption as adults. With sufficient input (Wonnacott et al. 2012), and/or sufficient scaffolding (Bencini & Valian 2008), children are of course ultimately capable of learning the forms and functions of abstract constructions (Tomasello 2003). But much previous work has found that younger children are less willing to produce novel verbs in unwitnessed constructions than are older children and adults (e.g., Akhtar 1999; Boyd & Goldberg 2009; Theakston 2004; Tomasello 2000, 2003). For this reason, we might expect children to show greater lexical conservativism than adults, possibly because children are not able to recognize the intended functions of abstract constructions as readily as adults, or because children may be more likely to try to imitate the adult experimenter as closely as possible. At the same time, we also know that children are at times more likely to generalize beyond their input on the basis of formal properties in order to simplify, particularly when the item-specific properties are (or are perceived to be) functionless (Hudson Kam & Newport 2005). This may be due to a tendency to simplify input that is too complex to keep track of. Children also may require a great deal of input before they take advantage of the sort of indirect negative evidence that statistical preemption provides (Brooks & Tomasello 1999; Hao 2015; Goldberg & Boyd 2015).

Therefore, it is possible that young children will either show greater lexical conservatism than adults, or they may overgeneralize one construction without regard to abstract differences in interpretation. Further work is required to determine at what age children begin to show the same degree of generalization found for adults in Experiment

1, and at what age they are as responsive to evidence of statistical preemption as the adults in Experiment 2.

5. *Conclusion*

We have seen that adult learners are exquisitely sensitive to the form and function of novel constructions, and to the distribution of verbs in terms of both their formal properties and their contexts of use. In particular, results from the first experiment demonstrate that speakers readily learn the functions associated with two distinct novel constructions and spontaneously generalize beyond the input on the basis of these learned functions. Stable verb-construction mappings in the input were largely ignored as speakers selected whichever of the two phrasal constructions better suited their intended message.

Similar previous findings had suggested that learners tended to use constructions with new verbs because particular semantic scenes were better predictors of constructions than were verbs (Thothathiri and Rattinger 2016), but in the current Experiment 1, verbs and semantic scenes were both perfect predictors of the choice of construction used, but learners strongly favored preserving the scene-construction mapping rather than the verb-construction mapping they had witnessed during exposure. We suggest that this preference stems from the communicative advantage of making use of constructional meaning (scene-construction mapping). Future work is required to determine whether children distinguish constructions on the basis of function as readily as adults do.

We did see a small but significant effect in a comparison of the lexicalist and alternating conditions in Experiment 1 that can naturally be attributed to a difference in cue reliability. In particular, when two out of six of the verbs alternated, participants were even more likely to preserve the constructional meaning at the expense of verb-construction distributions witnessed in the input, apparently because verbs were less reliable cues to which construction should be used than were the type of scenes. We also saw a possible effect of cue-reliability in Experiment 2, where the degree of affectedness played a markedly reduced role in production and judgment data than in Experiment 1, and where the reliability of the degree of affectedness to predict which construction would be used was reduced. The results of Experiment 2 also serve to reassure us that learners are capable of learning verb-specific distributions with the amount and type of input provided. Moreover, results demonstrate some effect of context for the non-preempted verbs and particularly for novel verbs in Experiment 2, which demonstrates that participants were simultaneously sensitive to the functions of the learned constructions.

Mini-artificial language learning studies often raise thorny issues about what exactly is learned in the experimental context and what is an effect of prior knowledge of the natural language already spoken by the participants (Fedzechkina, Jaeger, &

Trueswell, 2015; Goldberg 2013; Willits, Amato, & MacDonald 2015). The present experiments aimed to reduce the effect of prior knowledge of English in three ways. First the novel constructions involved both non-English word orders and abstract meanings not associated with English word order constructions. Moreover, unlike generalizations found in previous work (Perek & Goldberg 2015**;** Thothathiri & Rattinger 2016), the present tendency to generalize on the basis of the constructions in Experiment 1 is not easily attributable to prior knowledge about the balance between verbal semantics on the one hand, and information structure properties or adjunct status on the other. Finally, Experiment 2 allows us to rule out the possibility that the efficacy of statistical preemption necessarily relies on prior knowledge that a particular construction happens to be constrained in lexically idiosyncratic ways.

Thus, the key contributions of the present paper include a clear demonstration that learners are capable of generalizing on the basis of the learned semantics associated with two distinct abstract constructions without reliance on relevant prior knowledge (Experiment 1), while avoiding overgeneralizations when there is evidence that a verb is statistically preempted from occurring in one of the two constructions (Experiment 2).

Acknowledgements

We are grateful to Isaac Treves for running participants.

References

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

Ambridge, B. & Goldberg, A. E. (2008). The island status of clausal complements: evidence in favor of an information structure explanation. *Cognitive Linguistics, 19*(3), 349-381.

Ambridge, B., Pine, J. M., Rowland, C. F., Freudenthal, D., & Chang, F. (2014). Avoiding dative overgeneralisation errors: semantics, statistics or both?. *Language, Cognition and Neuroscience*, *29*(2), 218-243.

Ambridge, B., Pine, J. M., Rowland, C. F., & Chang, F. (2012). The roles of verb semantics, entrenchment and morphophonology in the retreat from dative argument structure overgeneralization errors. *Language, 88*(1), 1–60.

Ambridge, B., Pine, J. M., Rowland, C. F., & Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children’s and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, *106*(1), 87–129.

Amato, M. S., & MacDonald, M. C. (2010). Sentence processing in an artificial language: Learning and using combinatorial constraints. *Cognition*, *116*(1), 143–8. http://doi.org/10.1016/j.cognition.2010.04.001

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*, 67-82.

Aronoff, M. (1976). Word formation in generative grammar. *Linguistic Inquiry* Monographs Cambridge, Mass., (1), 1-134.

Arunachalam, S., & Waxman, S. R. (2014). Let's See a Boy and a Balloon: Argument Labels and Syntactic Frame in Verb Learning. *Language Acquisition*, (just-accepted).

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry, 10*, 533–581.

Bannard, C. & Matthews, D. (2008). Stored Word Sequences in Language Learning: The Effect of Familiarity on Children's Repetition of Four-Word Combinations. *Psychological Science, 19*(3), 241-248.

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic* (Vol. 8). John Benjamins Publishing.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673-721.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bates, D., Maechler, M., Bolker, B. & Walker, S. 2011. *lme4: Linear mixed- effects models using S4 classes*. R package. URL: http://CRAN.R-project.org/package=lme4

Bencini, G. M., & Valian, V. V. (2008). Abstract sentence representations in 3-year-olds: Evidence from language production and comprehension. *Journal of Memory and Language*, *59*(1), 97-113.

Bod, R. (1998). *Beyond Grammar: An Experience-Based Theory of Language*. CSLI Publications, Stanford.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In J. Hawkins (Ed.), *Explaining language universals* (pp. 73-101). Oxford: Basil Blackwell.

Bolinger, D. (1968). Entailment and the Meaning of Structures. *Glossa, 2*, 119-127.

Boyd, J. K. & Goldberg, A. E. (2011). Learning what not to say: the role of statistical preemption and categorization in "a"-adjective production. *Language, 81*(1), 1- 29.

Boyd, J. K. and Goldberg, A. E. (2012). Young children fail to fully generalize a novel argument structure construction when exposed to the same input as older learners. *Journal of Child Language*, *39*, 457-481.

Boyd, J. K., Gottschalk, E. A., & Goldberg, A. E. (2009). Linking rule acquisition in novel phrasal constructions. *Language Learning*, *59*(s1), 64-89.

Braine, M. D. (1971). On two types of models of the internalization of grammars. In Dan I. Slobin (Ed.), *The ontogenesis of grammar: a theoretical symposium*. New York, NY: Academic Press.

Braine, M. D. S., Brody, R. E., Brooks, P., Sudhalter, V., Ross, J. A., Catalano, L., & Fisch, S. M. (1990). Exploring Language Acquisition in Children with a Miniature Artificial Language: Effects of Item and Pattern Frequency, Arbitrary Subclasses, and Correction. *Journal of Memory and Language*, *29*, 591–610.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. *Cognitive foundations of interpretation*, 69-94.

Bresnan, J. & Ford, M. (2010). Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. *Language* 86(1): 186-213.

Brooks, P. J., & Tomasello, M. (1999). How children constrain their argument structure constructions. *Language*, *75*(4), 720–738.

Bybee, J. (1985). *Morphology: a study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambrige University Press.

Bybee, J. L., & Hopper, P. J. (Eds.). (2001). *Frequency and the emergence of linguistic structure* (Vol. 45). John Benjamins Publishing.

Bybee, J., & Thompson, S. A. (1997). Three Frequency Effects in Syntax. *Berkeley Linguistic Society*, *23*.

Casenhiser, D. & Goldberg, A. E. (2005). Fast Mapping of a Phrasal Form and Meaning. *Developmental Science*, 8, 500-508.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale,NJ: Lawrence Erlbaum.

Culbertson, J. and Elissa Newport. (2015) Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71-82.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122, 306-329.

Dąbrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, *16*(3), 437-474.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in second language acquisition*, *24*(02), 143-188.

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*(3), 254-260.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National*

Academy of Sciences, *109*(44), 17897-17902.
Fedzechkina, M., Jaeger, T. F., & Trueswell, J. C. (2015). Production is biased to provide informative cues early: Evidence from miniature artificial languages.
Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, *54*(4), 541-553.
Fisher, C. (2002). Structural limits on verb mapping: the role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science*, *5*(1), 55-64.
Fisher, C., Gleitman, H., & Gleitman, L. R. (1991). On the semantic content of subcategorization frames. *Cognitive psychology*, *23*(3), 331-392.
Ford, M., Bresnan, J., & Kaplan, R. M. (1982). A competence based theory of syntactic closure. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp. 727–796). Cambridge, MA: MIT Press.
Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 748-775.
Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37, 58–93.
Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language processing*. Psychology Press.
Gibson, E., Schutze, C. T., & Salomon, A. D. A.-J. (1996). The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research*, *25*(1), 59–92
Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
Goldberg, A. E . (2004). "Pragmatics and Argument Structure." *The handbook of pragmatics*. L. Horn, & G. Ward (eds.). John Wiley & Sons.
Goldberg, A. E. (2011). Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics, 22*(1), 131-154.
Goldberg, A. E. (2013). Substantive learning bias or familiarity effect? Comment on Culbertson, Legendre and Smolensky (2012). Cognition 127 (3) 420-426
Goldberg, A. E., Casenhiser, D., & Sethuraman, N. (2004). Learning Argument Structure Generalizations. *Cognitive Linguistics, 15*, 289-316.
Goldberg, A. E., & Boyd, J. K. (2015). A-adjectives, statistical preemption, and the evidence: Reply to Yang (2015). *Language*, *91*(4), e184-e197.
Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in cognitive sciences*, *4*(5), 178-186.

Green, G. (1974). *Semantics and syntactic regularity*. Bloomington: Indiana University Press.

Gries, S. T., & Divjak, D. (Eds.). (2012). *Frequency effects in language learning and processing* (Vol. 1). Walter de Gruyter.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The Learnability and Acquisition of the Dative Alternation in English. *Language*, *65*(2), 203–257.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford University Press.

Hawkins, J. A. (2014). *Cross-linguistic Variation and Efficiency*. Oxford University Press.

Hovav Rappaport, M., & Levin, B. (2008). The English dative alternation: The case for verb sensitivity. *Journal of Linguistics*, *44*(01), 129–167.

Hao, J. (2015). Abstraction versus restriction in syntactic learning: An examination of children's acquisition of the a-adjective restriction. Princeton, NJ: Princeton University senior thesis.

Hudson Kam, C. & E. Newport (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* 1, 151 –195.

Jaeger, F. T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23-62.

Kiparsky, P. (1982). Word-formation and the lexicon. Mid-America Linguistics Conference.

Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588-611.

Lakoff, G. (1970). *Irregularity in Syntax*. New York: Holt, Rinehart and Winston.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. University of Chicago Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Langacker, R. (1987). *Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites*. Stanford University Press.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, *19*, 849.

Lewis, J. D., & Elman, J. L. (2000). Learnability and the Statistical Structure of Language:

Poverty of Stimulus Arguments Revisited. In *Annual Boston University Conference on Language Development*. Boston University.
Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208.
MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101, 676–703.
Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, 126, 313-318.
Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2005). The role of frequency in the acquisition of English word order. *Cognitive Development*, *20*(1), 121–136.
Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*(4), 393–424.
Moeser, S. D., & Bregman, A. S. (1972). The role of reference in the acquisition of a miniature artificial language. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 759-769.
Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R2 from Generalized Linear Mixed-effects Models. *Methods in Ecology and Evolution*, *4*, 133–142.
Oehrle, R. (1976). *The Grammatical Status of the English Dative Alternation*. PhD Thesis, MIT. Cambridge: The MIT Press.
Payne, J., Pullum, G. K., Scholz, B. C., & Berlage, E. (2013). Anaphoric *one* and its implications. *Language*, *89*(4), 794–829.
Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam: Benjamins.
Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(03), 607-642.
Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226-233.
Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122, 280-291.
Pierce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2):8-13.
Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, Mass: MIT Press/Bradford Books.
Pierrehumbert, J. (2001). Exemplar dynamics: word frequency, lenition and contrast.

Typological Studies in Language, *45*, 137–158.
Poser, William J. 1992. Blocking of phrasal constructions by lexical items. Lexical matters, ed. by Ivan Sag and Anna Szabolsci, 111–130. Stanford: CSLI.
Reali, F., & Christansen, M. H. (2005). Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. *Cognitive Science*, *29*(6), 1007–1028.
Rendell, L. (1986). A general framework for induction and a study of selective induction. *Machine Learning*, *1*(2), 177-226.
Robenalt, C. & Goldberg, A. E. (to appear). Judgment and frequency evidence for Statistical Preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics.*
Saffran, J. R. (2001). Words in a sea of sounds: the output of infant statistical learning. *Cognition*, *81*, 149–169.
Saffran, J. R. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.
SAS Institute Inc (1978). *SAS Technical Report R-101. Tests of Hypotheses in Fixed-Effects Linear Models*. Cary, NC: SAS Institute Inc.
Schumacher, R. A., Pierrehumbert, J. B., & LaShell, P. (2014). Reconciling Inconsistency in Encoded Morphological Distinctions in an Artificial Language Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci2014) Austin , TX : Cognitive Science Society.
Scott, R. M., & Fisher, C. (2009). Two-year-olds use distributional cues to interpret transitivity-alternating verbs. *Language and cognitive processes*, *24*(6), 777-803.
Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(33), 11629–34.
Suttle, L., & Goldberg, A. E. (2011). The partial productivity of constructions as induction. *Linguistics*, *49*(6), 1237-1269.
Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind*. Oxford University Press.
Theakston, A. L. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development*, *19*(1), 15-34.
Thothathiri, M., & Rattinger, M. G. (2016). Acquiring and Producing Sentences: Whether Learners Use Verb-Specific or Verb-General Information Depends on Cue Validity. *Frontiers in Psychology*, *7*.
Thompson, S. A. (1990). Information Flow and Dative Shift in English Discourse. In J. Edmondson, K. Feagin, & P. Mühlhäusler (Eds.), *Development and Diversity:*

Linguistic Variation across Time and Space. (pp. 239–253). Summer Institute of Linguistics.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*(3), 209-253.

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language*, *27*(1), 71-86.

Waxman, S. R., Lidz, J. L., Braun, I. E., & Lavin, T. (2009). Twenty four-month-old infants' interpretations of novel verbs and nouns in dynamic scenes. *Cognitive Psychology*, *59*(1), 67-95.

Willits, J. A., Amato, M. S., & MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cognitive psychology*, *78*, 1-27.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language* 65, 1–14.

Wonnacott, E., Boyd, J. K., Thompson, J. & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in five year olds. *Journal of Memory and Language, 66*, 458-478.

Wonnacott, E., E. Newport & M. Tanenhaus (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology* 56: 165-209.

Yang, C. (2015). Negative knowledge from positive evidence. *Language*, *91*(4), 938-953.

Xu, F. & Tenenbaum, J. B. (2007). Word Learning as Bayesian Inference. *Psychological Review* 114(2): 245.