Cognition 168 (2017) 276–293

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

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

Florent Perek [a,*], Adele E. Goldberg [b]

[a] University of Birmingham, Department of English Language and Applied Linguistics, 3 Elms Road, Edgbaston, Birmingham B15 2TT, United Kingdom
[b] Princeton University, Department of Psychology, Peretsman-Scully Hall, Princeton, NJ 08540, USA

ARTICLE INFO

Article history:
Received 7 August 2016
Revised 16 June 2017
Accepted 18 June 2017
Available online 27 July 2017

Keywords:
Language acquisition
Artificial language learning
Novel construction learning
Statistical learning
Argument structure constructions
Generalization

ABSTRACT

How do people learn to use language in creative but constrained ways? Experiment 1 investigates linguistic creativity by exposing adult 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 effect. 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 or to a difference in cue validity. In order to investigate how speakers learn to constrain generalizations, Experiment 2 includes one verb (out of 6) that was witnessed in a single construction to describe both strong *and* weak effects, essentially statistically 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.

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., Fisher, 2002; Gertner, Fisher, & Eisengart, 2006; Naigles, 2000).

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, 1988; Goldberg, 1995; Pinker, 1989). 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, 2000)
(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 long puzzled researchers (Ambridge, Pine, Rowland, & Chang, 2012; Baker, 1970; Bowerman, 1988; Braine, 1990; Goldberg, 1995; Lakoff, 1970; Perek, 2015; Pinker, 1989), and artificial language learning experiments have been found useful in addressing them (e.g., Braine et al., 1990; Brooks et al., 1993; Amato & MacDonald, 2010; Culbertson, Smolensky, & Legendre, 2012; Fedzechkina, Jaeger, & Newport, 2012; Gómez & Gerken, 2000; Moeser & Bregman, 1972; Valian & Coulson, 1988). A typical paradigm involves exposing learners to a miniature language which includes a set of novel word order patterns that are paired with familiar transitive or intransitive interpretations. Another paradigm involves exposing learners to novel *constructions* that pair novel word order patterns with *novel abstract meanings* (Casenhiser & Goldberg, 2005); speakers need to learn constructions in order to produce and comprehend real natural languages; i.e., they need

* Corresponding author.
E-mail addresses: f.b.perek@bham.ac.uk (F. Perek), adele@princeton.edu (A.E. Goldberg).

http://dx.doi.org/10.1016/j.cognition.2017.06.019
0010-0277/

knowledge of how words can be combined formally *and* the particular range of interpretations that each type of combination evokes (Goldberg, 2006; Tomasello, 2003).

One factor that plays a role in determining whether speakers are willing to generalize the way a verb is used is whether other verbs have already been witnessed being generalized. For example, Wonnacott, Newport, and Tanenhaus (2008) exposed adult participants to an artificial language that included two synonymous transitive constructions. Results demonstrated that participants are sensitive to the overall statistics of an artificial language when determining whether predicates can be extended in new ways. In particular, participants tended to behave conservatively when exposed to a language in which all 12 verbs appeared in only one of two constructions, i.e., they avoided extending verbs for use in the other construction (see also Perek & Goldberg, 2015, Exp. 2; Thothathiri & Rattinger, 2016, Exp.1). However, when exposed to a language in which some of the verbs were witnessed in both constructions, they showed some degree of generalization, using verbs freely in either construction. Wonnacott (2011) is a similar study that replicated the basic findings with children.

Note that when distinct formal patterns are assigned the exact same function, using a verb in one construction conveys exactly the same message as using a verb in the other construction. But in natural languages, it is hard to find verbs that occur in two constructions that serve exactly the same function; instead the choice between two constructions is typically conditioned by differences in information structure or semantics (e.g., Bolinger, 1968; Bresnan, 2011; Goldberg, 1995; Scott-Phillips, Kirby, & Ritchie, 2009). With this in mind, Perek and Goldberg (2015, Exp. 1) aimed to investigate whether communicative pressures would encourage learners to generalize the constructions for use with verbs that had not been witnessed in those constructions during exposure. Adult participants were exposed to six nonce verbs that were used in two constructions that differed in terms of information structure properties as well as word order. In particular, one construction always contained a pronominal patient argument ($\text{Pronoun}_{\text{Patient}}$ NP_{Agent} V), while the other occurred exclusively with lexical noun phrase arguments in a distinct order (NP_{Agent} $\text{NP}_{\text{Patient}}$ 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 fully 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, Exp. 2) 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. One construction had Verb-Agent-Patient order and included an additional, final nominal that was interpreted as an instrument, and the other construction had Verb-Patient-Agent order and included a final nominal that was interpreted as a modifier (something the patient was holding). Ten out of 12 verbs consistently appeared in one or the other construction, while two verbs alternated between the two constructions. As found in Perek and Goldberg (2015), speakers demonstrated a strong tendency to generalize on the basis of the functions of the constructions, using verbs in whichever construction better captured the intended message.

The striking tendency in these studies for participants to generalize beyond the verb-specific input when the constructions' functions were distinct is, however, subject to a potentially potent criticism. The tendency to ignore verb-specific distribution may have resulted from prior knowledge about the sorts of information that individual verbs normally convey. The constructions used by Perek and Goldberg (2015) differed in terms of information structure, and adult participants can be expected to be aware 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, i.e., 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.

Additionally, previous experiments offered distinct interpretations of *why* participants are likely to generalize beyond their input when two constructions are assigned distinct functions. As described above, while Perek and Goldberg (2015) suggested that participants' productive use of verbs in an unwitnessed construction results from the communicative pressure to express an intended message with whichever construction is better suited, Thothathiri and Rattinger (2016) interpreted their parallel findings in terms of an advantage of *cue validity* of verbs vs. scenes in predicting which construction was expressed during exposure (Bates & MacWhinney, 1989; Chan, Lieven, & Tomasello, 2009; Goldberg, Casenhiser, & Sethuraman, 2005; MacWhinney, 2012). In the latter experiment, the type of scene predicted which construction was used with a probability of 1. On the other hand, while 10 out of 12 verbs also uniquely predicted which construction was witnessed during exposure, offering a cue validity of 1, another two verbs appeared in either construction with equal probability, giving them a cue validity of 0.50. Thus, the cue validity across all verbs for predicting the construction was 0.92 ($= 1 \times 5/6 + 0.5 \times 1/6$). The authors conclude that learners used the scene rather than the verbs to determine which construction to use because the scenes were more reliable predictors of constructions than verbs.

Two experiments presented here aim to investigate how learners generalize beyond their exposure and how those generalizations are constrained. The experiments are also designed to address issues raised by previous work, namely: (a) the possible confound that prior knowledge of the division of labor between verbs and constructions led to an increase in generalization and (b) the question of whether cue validity or expressive power (or both) encourages the productive use of constructions. In both experiments, 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 a "patient" (or undergoer) argument; the other construction describes actions with a weaker but otherwise similar effect. This 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 phrasal 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 patient 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 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 four out of six verbs were witnessed consistently in one construction or the other, but two of the six verbs were witnessed in both constructions. Thus in the lexicalist condition, the verbs and scenes are both perfect predictors of the choice of construction witnessed during exposure: verbs predict which construction is used with a probability of 1 and the degree of effect on the patient also predicts the choice of construction with a probability of 1. Experiment 1 thus allows us to investigate whether or how learners generalize when cue validity for scenes and verbs are matched. In a second condition, an "alternating" condition, *scenes* again perfectly predict which construction is used during exposure, but 2/6 of the verbs occur in both constructions, rendering the overall cue validity of verbs equal to 0.83 (=1 × 4/6 + 0.5 × 2/6). Since the communicative demands are held constant across the lexicalist and alternating conditions, if the constructions display greater productivity in the alternating condition, it will be attributable to the reduced cue validity of verbs in that condition.

A second experiment investigates how generalizations are constrained, allowing speakers to avoid overgeneralizations (such as *??giggle me*). Experiment 2 exposes a separate group of adult participants to evidence that one of six verbs is *statistically preempted* from occurring in one of the same two novel constructions used in the first experiment. In particular, one verb is witnessed in only one construction *even when the semantics is congruent with the other construction*. In Experiment 2, the cue validity of verbs to predict which construction is used is 1, while the cue validity of scenes to predict constructions is 0.92 (1 verb out of 6 is used with scenes that are incongruent with the construction half of the time). The design of this experiment will allow us to investigate whether the reduced cue validity of scenes will lead to overall conservative behavior of verbs, since the verbs are better predictors of constructions than scenes. We hypothesize that adults will in this case constrain the distribution of this verb by avoiding using it productively in the unwitnessed construction. A comparison of Experiments 1 and 2 will moreover allow us to determine whether speakers are able to keep track of whether one verb occurs in both types of scenes (Exp. 2 only), or whether they are instead primarily interested in tracking the formal distribution of verbs (the lexicalist condition of Exp. 1 provides the same formal distribution as Exp. 2).

2. Experiment 1

In a between-subjects design, participants were assigned to either the lexicalist condition or the alternating condition. Each of two word order constructions witnessed during exposure was always appropriately used to describe an accompanying scene. That is, a "Weak-Cx" was witnessed when the effect on the patient was weak, and a "Strong-Cx" was witnessed when the effect on the patient was strong (see Materials for the formal properties of the two constructions). Thus, there were two conditions:

lexicalist condition: each participant witnessed 3 verbs only occurring in one word order construction ("Weak-Cx") accompanied by scenes in which unique actions were performed with weak effects on the patient argument; and 3 other verbs only occurring in a different word order construction ("Strong-Cx") accompanied by scenes in which unique actions were performed with strong effects on the patient argument.
alternating condition: each participant witnessed 2 verbs only occurring in one construction, 2 verbs only occurring in the other construction, and 2 verbs occurring in each construction 50% of the time.

The differences between the two conditions are summarized in Fig. 1. In both conditions, all uses of each construction were congruent semantically: scenes displaying weak effects on the patient argument were matched with descriptions using one word order, and scenes displaying strong effects were matched with descriptions using the other word order. At test, participants were asked to describe similar scenes that involved either a strong or a weak effect on a patient argument.

If speakers base their productions solely on the basis of distributional evidence in the input, we would expect speakers to restrict their productions, using each verb only in the construction in which it had been witnessed. If, however, speakers prefer to use constructions that match the scene, 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 and Goldberg's (2015, Exp. 1) 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

24 undergraduate students at Princeton University took part in the study. 18 of them participated in the experiment for course credit, and the other six received payment. 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: Patient Agent Verb (PAV) or Agent Patient Verb (APV). A suffix *-po* was appended to the patient argument in order to disambiguate between the two word orders (e.g., *the cat-po*). Each of six verbs had a distinct meaning including: blow-on (the agent bends over and blows air at the patient), headbutt, kick, punch, push, slap (with both hands), spin (the agent spins towards and hits the patient), swirl-strike (the agent strikes the patient with a swirling blow).

The word order and semantics of the two constructions were distinct: one word order construction (Agent-Patient-Verb) hereafter the **Strong-Cx,** always described actions that had a strong effect on the patient argument (4), during exposure; a second word order construction (Patient-Agent-Verb), hereafter the **Weak-Cx** always described actions that had weak effects (5):

(4) NP_{Agent} $[NP\text{-}po]_{Patient}$ V (APV order, hereafter, **Strong-Cx**)
"agent acts on patient causing a strong effect"
e.g., *the rabbit the cat-po mooped*

(5) $[NP\text{-}po]_{Patient}$ NP_{Agent} V (PAV word order, hereafter, **Weak-Cx**)
"agent acts on patient causing a weak effect"
e.g., *the panda-po the pig pilked*

Both versions of each action (strong effect and weak effect) were enacted by anthropomorphized animals in 3D animations recorded as video clips.[1] Strong effects consisted in, for example, a patient (an animal figure) moving rapidly all the way across and off the screen

[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.

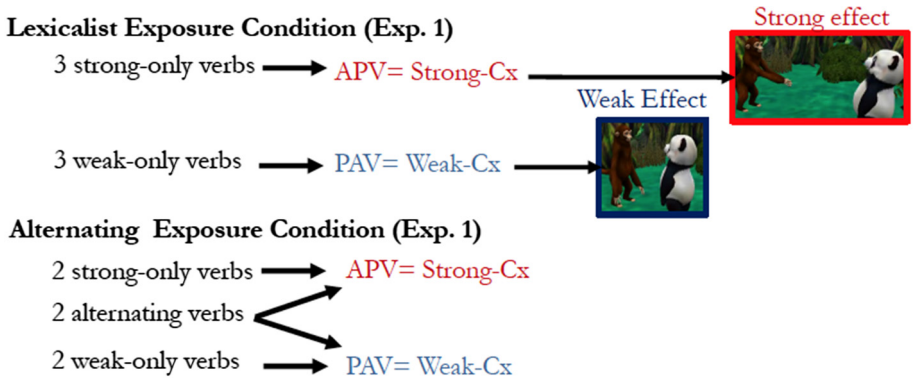


Fig. 1. The two types of exposure provided in the Lexicalist and Alternating conditions in Experiment 1. (The color-coding is only included for expository purposes with blue corresponding to "weak" and red corresponding to "strong".).

while performing dramatic gestures like throwing her arms backwards, arching her back at a 90 degree angle, etc. Weak effects involved the patient moving only slightly and performing similar but less ample gestures. At the end of strong-effect scenes, the patient 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. The distinction between weak and strong effect is illustrated by Fig. 2A and B.

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, after exposure, participants were asked to produce sentences to describe scenes that involved either strong or weak effects on the patient 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 instructions were given in written form on the computer screen. For each participant, the experiment was conducted over two sessions, between 24 and 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 label of the animal: "this is the panda/rabbit/etc." They were then exposed to six verbs 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 of 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.

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 patient argument 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: 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 patient, and twice with a scene showing a strong effect, each time with a different pair of agent and patient arguments. In all tasks, the left-to-right orientation of the patient 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

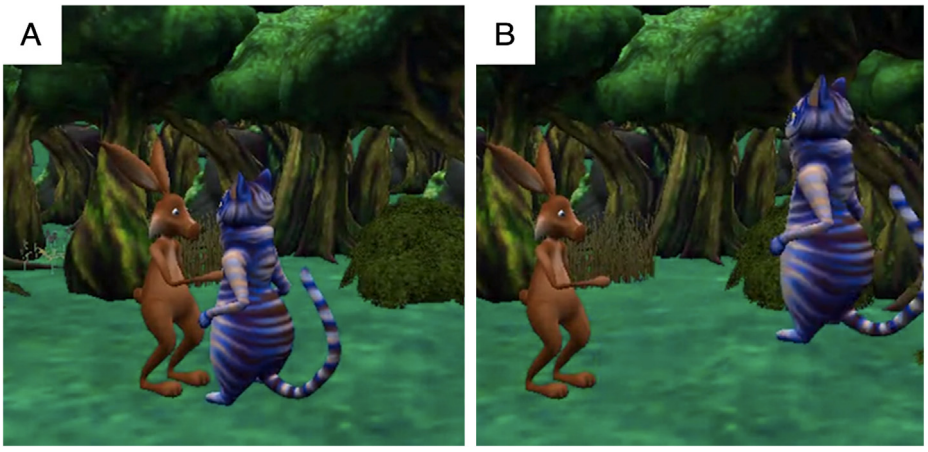


Fig. 2. (A) Sample screen shot of a video showing a **weak effect** on patient: a rabbit punches a cat, with weak effect on the cat. (B) Corresponding screen shot of a video showing a **strong effect** on the patient: same action produces a strong effect on the cat.

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 triple of tasks is illustrated by screenshots in Fig. 3.

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 patient 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 consisted of 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 Fig. 4.

Participants provided responses on a 7-point Likert scale, with 1 being "sounds bad" and 7 being "sounds good."[2] All six verbs that had been witnessed during the exposure phase were used four times each, once in each of the following combinations of sentence and type of scene: Congruent combinations involved the Strong-Cx with a strong effect on the patient, and Weak-Cx with a weak effect; Incongruent combinations involved the Weak-Cx with a strong effect on the patient, and Strong-Cx 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 whether the meaning of the sentence matched the scene shown to them.

2.4. Results

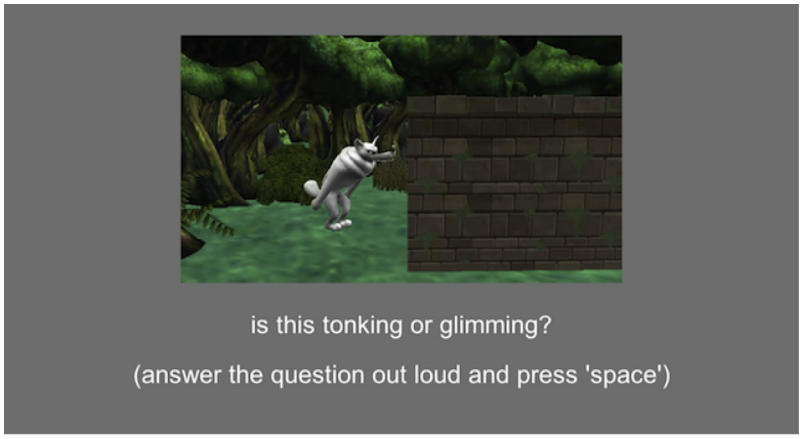
Because we are interested in language use and not language learning per se, we focus below on the data collected after the second and final day of exposure, i.e., at the outcome of the learning process. We describe the results of the production and sentence rating tasks in turn.[3] Our entire dataset (including the data from both day 1 and day 2) is available as an online supplement.

2.4.1. Production task

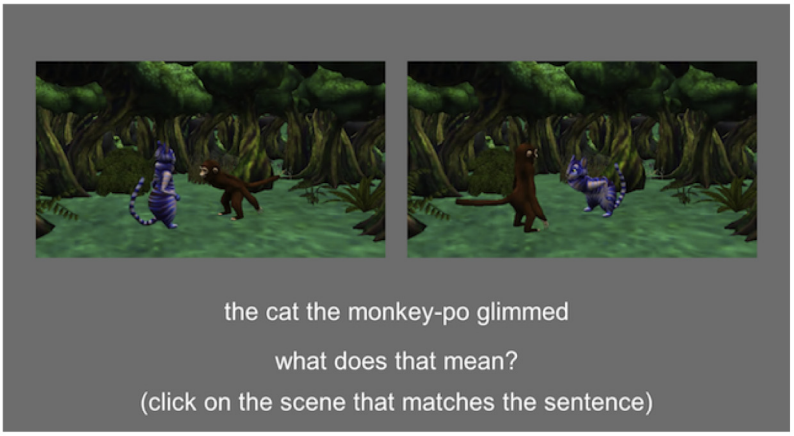
The results of the production task were coded according to which word order was used. Sentences consisting of a regular noun phrase referring to the agent, a noun phrase followed by the particle *–po* referring to the patient, and the verb (in that order: APV), were coded as instances of the Strong-Cx, regardless of whether the scene actually involved a strong effect. That is, **the coding of construction was determined by word order only, not the semantics of scenes**. Sentences consisting of the same noun phrases in the opposite order (patient then agent) followed by the verb (PAV) were coded as instances of the Weak-Cx. 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, including cases in which participants used the right order of arguments but attached the particle *–po* to the wrong noun phrase (i.e., the agent). 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. 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). When the subject hesitated or produced multiple sentences, only their last full production was considered. Even though the correct

[2] One of the reviewers points out that this task did not provide participants with an option to signal that they did not know the answer, and that in such cases they could have defaulted to the middle value ('4'), thus potentially biasing responses towards this value. We examined the datapoints for the '4' ratings as suggested by the reviewer, but we did not find anything unusual about them; importantly, it is not an unusually frequent rating given by participants. We used logistic regression to test whether any of the predictors and their interactions positively influenced the choice of the '4' rating (as opposed to the other six), but we did not find any significant trends.

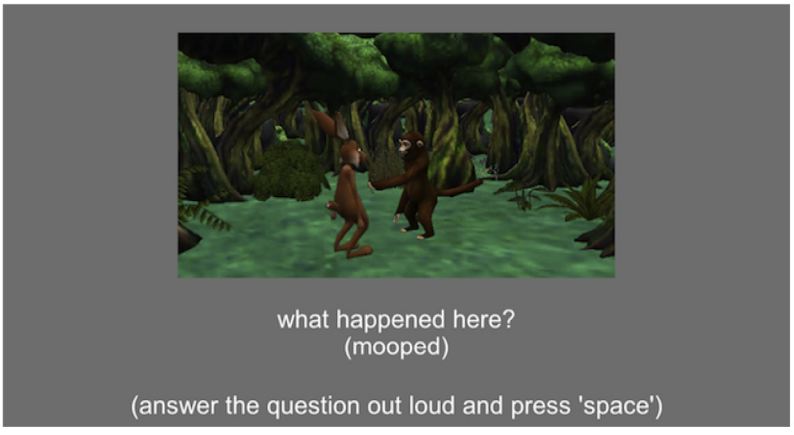
[3] As intended, by day 2, performance on the comprehension task was at ceiling in that participants were successfully able to assign thematic roles to the arguments of the verbs, identifying the correct scenes 98.9% of the time.



Vocabulary distractor task



Comprehension distractor task



Production task

Fig. 3. Screenshots of the three tasks given in each comprehension/production test triple. Testing consisted of 32 such triples.

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 Weak-Cx and Strong-Cx productions are plotted in Fig. 5, separately for the lexicalist and alternating conditions. The same general trend is found for all verb types, regardless of how the verb was witnessed during exposure. The learned constructions were generally used appropriately in both conditions, with results confirming that this was maximally the case (only) in the alternating condition. The Weak-Cx construction tended to be used when the effect on the patient was weak, and the Strong-Cx construction tended to be used when the effect on the patient was strong.

To test for statistical significance, we submitted the data to mixed effects logistic regression, using the package lme4 in the R

Fig. 4. Example screenshot of the sentence rating task, with the verb *wub* used in the Strong-Cx.

environment (Bates, Maechler, Bolker, & Walker, 2011).[4] Each production of the Weak-Cx or Strong-Cx is one observation in the dataset. The dependent variable, Strong-Cx (binary), records whether the utterance used the word order associated with the Strong-Cx vs. the Weak-Cx. 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 each construction during exposure; or whether participants used the constructions 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 patient: a binary variable that captures whether the scene involved a strong or weak effect on the patient (strong vs. weak);
(b) VerbType: a categorical variable that captures whether a verb had been witnessed during exposure only in the Strong-Cx construction (strong-only), only in the Weak-Cx construction (weak-only), or not witnessed at all in the input (novel).
(c) 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).

For this and all subsequent models, we followed an automatic stepwise model selection procedure (Baayen, 2008), whereby the most complex model containing all interactions between fixed effects was first fitted to the data and then compared, by means of likelihood ratio tests, to simpler versions of the same model where one effect is removed, in order to estimate whether this effect makes a significant contribution to the model, or whether it can be dispensed with without losing predictive power.[5] The results of the likelihood ratio tests for the full model can be found in Appendix A. The final model contains Condition, Effect, and VerbType as main effects, and the interaction between Condition and Effect. The fixed effects of this model are reported in Table 1. In order to appropriately measure main effects and interactions, sum contrast was used for all factors in the model, which means that the effect of all factors and interactions is measured with respect to the overall mean of the dependent variable, and not with reference to a baseline level of each variable (as in treatment contrast, commonly used by default in logistic regression). As Condition and Effect are binary variables, we only report the effect of one level ("alternating condition" and "strong effect" respectively), since, with sum contrast, the other level is defined to have an opposite effect of the same magnitude; the same applies to the interaction of these factors. The effects of all three levels of VerbType are reported individually.

Random effects for subjects (Subject), verb forms (VerbForm), and verb meanings such as whether the verb referred to a KICK or a SPIN (VerbMeaning) 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 associated with particular verb forms or meanings. We followed Barr, Levy, Scheepers, and Tily (2013) in starting with a maximal random effect structure containing random intercepts for Subject, VerbForm, and VerbMeaning, and by-participant random slopes for the factors Effect and VerbType. The model initially failed to converge, and only did so when we removed all random slopes and the random intercepts for VerbForm and VerbMeaning, thus only keeping random intercepts for Subject ($SD = 0.3475$). The variance of VerbForm and VerbMeaning is extremely small (below 0.0001), which means that these factors have very little effect on the subjects' productions, and that including them in the model would not make a noticeable difference. The same random effect structure was used for all models reported in this paper. Classification accuracy (i.e., the percentage of data points for which the model predicts the right construction) was 78.47%, and the area under the ROC curve (AUC) was estimated at 84.94%,[6] indicating that the model is a good fit for the data.

Since uses of the Strong-Cx 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 Strong-Cx, and conversely, negative values indicate that the factor favors the use of the Weak-Cx.

As was evident in Fig. 5, which construction a verb was used with depended strongly on the semantics of the scenes involved: i.e., whether the patient 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 Strong-Cx is significantly more likely to be used in the presence of a strong effect on the patient. In fact, Effect had a significant impact on how a verb was used, regardless

[4] We used the 1.1–12 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 and Schielzeth's (2013) method, as implemented by the MuMIn package version 1.15.6.

[5] We used the mixed function in the R package afex to perform this procedure automatically.

[6] For this and all subsequent logistic regression models, the classification accuracy and the AUC measure are reported in the caption of the relevant table. The AUC scores were calculated using the auc function from the R package pROC.

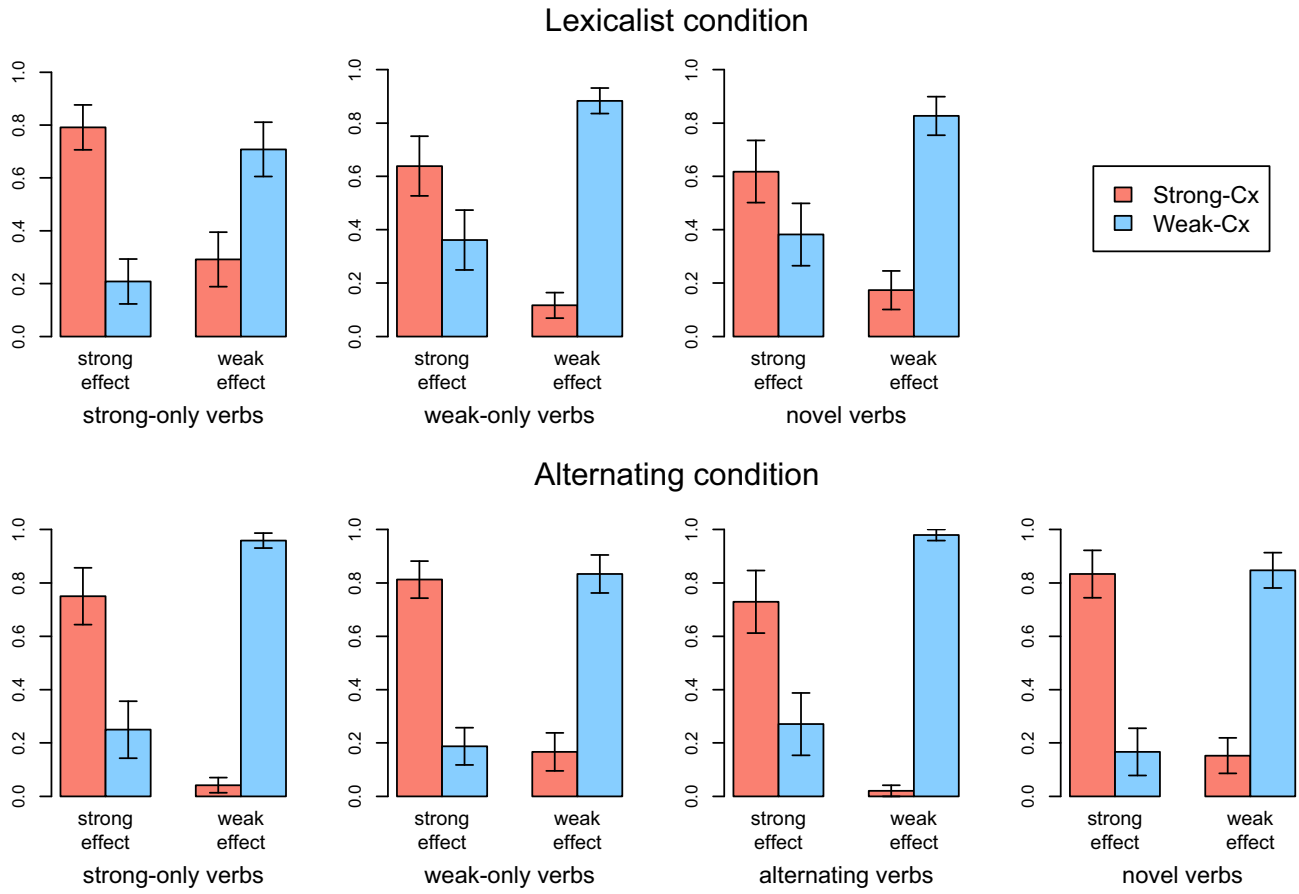


Fig. 5. Results from Experiment 1. Each of the seven panels includes the proportion of participants' productions that describe scenes in which there was a strong effect on the patient (left side) or a weak effect on the patient (right side). Proportions of Weak-Cx and Strong-Cx productions for the strong-only verbs presented only in the Strong-Cx construction, weak-only verbs presented only in the Weak-Cx construction, and for new novel verbs. The Alternating condition included two verbs that appeared in both constructions with congruent semantics; performance on these is represented in the third panel on the bottom row.

of how that verb had been witnessed in the input; that is, the interaction of Effect with VerbType was not significant when included in the model (see Appendix A).[7]

At the same time, the significant interaction of Condition and Effect indicates that the impact of Effect was more pronounced in the alternating condition than in the lexicalist condition (since the estimate of the interaction effect is positive). There are also significant though much weaker effects of VerbType: strong-only verbs were more likely to be produced with the Strong-Cx, and weak-only verbs more likely to be produced in the Weak-Cx construction. Novel verbs only had a marginally significant negative effect on Strong-Cx production, indicating that they were used much like the other verbs in participants' productions.

In sum, there is some evidence for an effect of lexical conservatism, but it is weak compared to that of constructional meaning. 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 patient was strong or weak. This was especially true in the alternating condition. Whether verbs had only been witnessed in the Strong-Cx or the Weak-Cx during exposure had comparatively little impact on their choice of construction, even in the lexicalist 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 with a value that indicates by how many standard deviations it diverges from the subject's average rating.[8]

Fig. 6 presents the distributions of z-scores in the two input conditions in the form of boxplots. The distributions are grouped by how the verb was witnessed during exposure: strong-only, weak-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 patient found in the stimuli set, from left to right: Strong-Cx with strong effect, Weak-Cx with weak effect, Strong-Cx with weak effect, and Weak-Cx with strong effect. The first two (in green) are combinations attested in the input: these are "effect-congruent", in that the scene described by the sentence matches observed usage of the construction in terms of whether the effect on the patient was strong or weak. The latter two (in red) conflict with the input and are therefore "effect-incongruent".

[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).

[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, was 0. The final dataset analyzed in this section totals 552 observations.

Table 1
Experiment 1 comparison of lexicalist and (partially) alternating condition on production task. Fixed effects of the logistic regression model predicting the production of the Strong-Cx construction. Classification accuracy = 78.47%, *AUC* = 84.94%. Marginal R^2 = 40.93%, Conditional R^2 = 43.02%., Model formula: StrongCx ~ Condition * Effect + VerbType + (1 | Subject).

	Estimate	Std. error	z-value	p-value
(Intercept)	−0.0009	0.1271	−0.007	0.9945
Condition (alternating)	0.3449**	0.1272	2.711	0.0067
Effect (strong)	1.4600***	0.1114	13.108	<0.0001
VerbType (strong-only)	0.5529***	0.1401	3.945	<0.0001
VerbType (weak-only)	−0.2947*	0.1390	−2.120	0.0340
VerbType (novel)	−0.2582	0.1483	−1.741	0.0817
Condition (alternating) × Effect (strong)	0.2778**	0.1056	2.630	0.0085

The stars next to the estimates reflect the significance threshold of the p-value: *** for p < 0.001, ** for p < 0.01, * for p < 0.05.

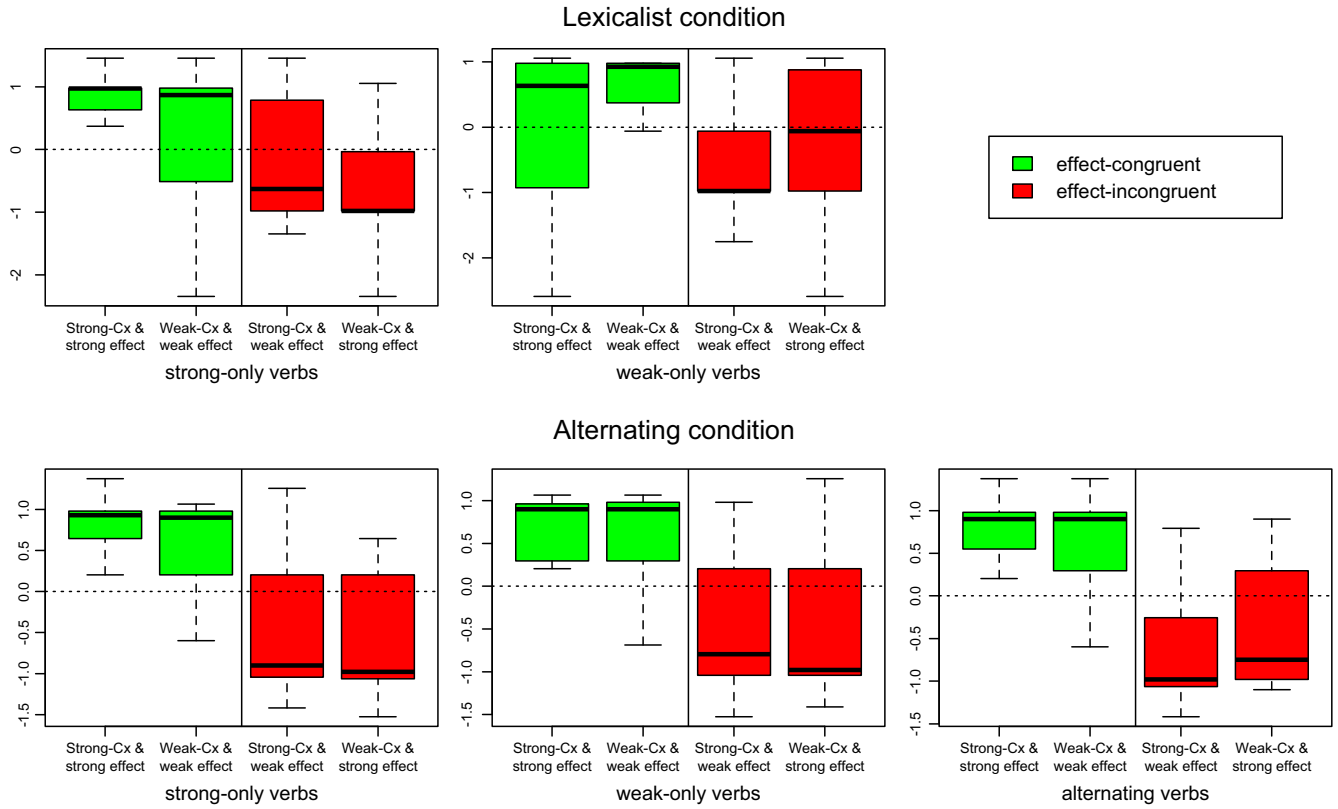


Fig. 6. Experiment 1 sentence ratings. Box plots of the distribution of grammaticality ratings (z-scores) in the lexicalist condition (top) and in the condition in which one third of verbs alternated (bottom), for each verb type, and each combination of construction and effect on the patient seen in the stimuli. Combinations that were congruent with the input regarding the effect on the patient (i.e., Strong-Cx with strong effect, Weak-Cx with weak effect) are plotted on the left-hand side of each box (in green); the other, incongruent combinations are plotted on the right-hand side of each box and colored (in red). Outliers are not plotted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As is standard in boxplots, 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 interquartile range (i.e., the difference between the upper and lower quartiles).

As can be seen in Fig. 6, participants in both the lexicalist and the alternating condition generally judged as more acceptable constructions that described congruent scenes: the Strong-Cx when the effect on the patient was strong, or the Weak-Cx 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 extend 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 Strong-Cx was used when the effect is

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

	Estimate	Std. error	t-value	p-value
(Intercept)	−0.0071	0.0368	−0.192	0.8479
EffectCongruent (true)	0.4871***	0.0357	13.657	<0.0001
VerbCongruent (true)	0.1302***	0.0368	0.775	0.0004
Condition (alternating)	−0.0071	0.0368	−0.192	0.8479
EffectCongruent (true) × Condition (alternating)	0.0734*	0.0357	2.058	0.0401
VerbCongruent (true) × Condition (alternating)	−0.0878 *	0.0368	−2.385	0.0174

The stars next to the estimates reflect the significance threshold of the p-value: *** for p < 0.001, ** for p < 0.01, * for p < 0.05.

strong, or if the Weak-Cx is used when the effect is weak, and false otherwise. VerbCongruent is set as true for strong-only verbs used in the Strong-Cx, for weak-only verbs used in the Weak-Cx construction, and for alternating verbs used in either construction (in the alternating condition only), and it is set as false for strong-only verbs used in the Weak-Cx, and weak-only verbs used in the Strong-Cx. As previously, sum contrast was used for all variables. By-subject random intercepts were also included in the regression, although they did not capture significant variance (because z-scores were used) (SD < 0.0001).

The fixed effects of the regression model are reported in Table 2; the full model with all interactions can be found in Appendix B. In both conditions, participants rated the sentence more favorably if the construction matched the effect on the patient or, to a lesser extent, if the verb was used in a construction with which it was attested in the input; i.e., we find a positive main effect of EffectCongruent, as well as a positive, yet weaker, main effect of VerbCongruent. However, both predictors are involved in significant interactions with Condition: in the alternating condition, the effect of EffectCongruent is slightly stronger, and that of VerbCongruent slightly weaker; in other words, participants in the alternating condition had a higher tendency to assume that the effect on the patient was a critical factor when rating instances of each construction than participants in the lexicalist condition, and conversely, a lower tendency to base their grammaticality judgments on the basis of observed usage of the verb. Yet, both interaction effects are quite weak compared to the main effect of EffectCongruent, showing that, by and large, the main factor influencing grammaticality judgments in both conditions is the effect on the patient.

To summarize, the results of the sentence rating task are consistent with those of the production task. In both conditions, sentences were judged markedly more acceptable if the construction was compatible with the effect seen on the patient in the corresponding scene. In the lexicalist condition, sentences were judged to be somewhat less grammatical when they contained a verb used in a different construction from the one it had been attested with in the input. In the condition in which one third of the verbs are witnessed alternating, all verbs were judged equally grammatical in either learned construction, as long as the effect on the patient was appropriately either strong or weak.

2.5. Discussion

Results of Experiment 1 demonstrate that learners generalize on the basis of the constructions' meanings even if it requires ignoring constraints on how verbs were witnessed being used in the input. The design addresses a possible concern raised by previous work that had similarly found a willingness to ignore verb-specific patterns in the input as the earlier results may have relied on prior knowledge that may have predisposed participants to assume that the functional distinctions were not naturally attributed to individual verbs (Perek & Goldberg, 2015; Thothathiri & Rattinger, 2016). The present experiment varied whether an effect on the patient 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. 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 virtually 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 compare two possible interpretations of why participants are likely to generalize beyond their input when two constructions are assigned distinct functions. One possibility is that learners prefer to select the construction which better suits the discourse context and therefore affords more expressive power (Perek & Goldberg, 2015). Another possibility is that leaners rely on whichever cue, verb or scene, is a more reliable predictor the choice of construction (Thothathiri & Rattinger, 2016). We find evidence of both factors in the present results.

Recall that in the present lexicalist condition, the cue validities of verbs and scenes are matched: verbs predicted which construction was used with a probability of 1, and the degree of effect on the patient also predicted which construction was used with a probability of 1. However, instead of relying equally or randomly on verbs or 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 and 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 evidence in favor of cue validity as an additional factor. 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, as two of the six verbs were witnessed in both constructions which made the verbs less reliable cues. This was again also true in a comparison of Perek and Goldberg (2015)'s lexicalist and alternating conditions. Interestingly, the relevant cue validity is not determined by individual verbs, but by the statistics of the language overall (see also Wonnacott et al., 2008); 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 verbs' distribution and freely used and accepted whichever construction better matched the scene at test. To summarize, the cue validity across all verbs and scenes seems to play some role in how participants selected the appropriate construction, but the effect was relatively weak, as they displayed a tendency to ignore verb-specific behavior, even when verbs were perfect predictors of which construction to use. When cue validity is matched, speakers rely more on the scenes, overriding verb-specific functionless distributions in order to gain more expressive power by using whichever construction is better suited to describe the scene.

One might argue that participants tended to generalize on the basis of the constructions because it is just easier to learn two constructions and their corresponding functions and ignore the specific distribution of six verbs, than it is to learn the specific distribution of six verbs and ignore the functions of the two constructions. Therefore, perhaps participants simply failed to learn which construction was associated with which verb in the input. This possible explanation seems unlikely, given that previous work has found that participants are capable of learning, and inclined to use and accept, the verb-specific distribution of six verbs with the present amount of exposure, when there is no functional distinction between the two constructions (Perek & Goldberg, 2015; Wonnacott et al., 2008; Exp.2; Thothathiri & Rattinger, 2016, Exp. 1). Is it possible that the assignment of distinct functions to the two constructions interferes with participants' ability to learn the distribution of individual verbs? We will see that this issue is addressed by the results of Experiment 2, which we now turn to.

The demonstration that learners display a strong tendency to generalize on the basis of the functions of newly learned constructions raises the question as to how generalizations are constrained, since it is clear that speakers do not extend real verbs for use in familiar constructions willy-nilly. That is, even productive constructions often have lexical exceptions (Baker, 1979; Braine, 1971; Levin, 1993; Pinker, 1989). For example, it is well-known that although the English double-object construction is productively extended to new verbs (e.g., *I'll message you the link*), certain other verbs resist occurring in it, even though the meaning would be perfectly clear, and even when the information structure properties would be appropriate. For example, 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., Goldberg, 1995; Levin, 1993; Pinker, 1989; see Ambridge, Pine, Rowland, Freudenthal, & Chang, 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 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). This would mean 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 word forms, e.g., *feet* is used instead of *foots*. We know that the word, *feet,* is learned because learners systematically witness *feet* in contexts that would otherwise be appropriate for *foots* (Aronoff, 1976; Kiparsky, 1982).

Previous experimental evidence supporting the idea that statistical preemption explains the sorts of ill-formed sentences in (6) has involved production or judgment tasks with familiar English constructions (Boyd & Goldberg, 2011; Brooks & Tomasello, 1999; Robenalt & Goldberg 2015). For example, Brooks and Tomasello (1999) found that novel verbs witnessed intransitively (*It meeked*) were preempted from being used transitively (*She meeked it*) when a periphrastic causative was witnessed (*She made it meek*). But the use of familiar constructions raises the possibility that learners brought with them prior knowledge that these particular constructions were finicky about which predicates could occur in them. While the constructions' lack of full productivity may itself have been learned via statistical preemption as the earlier studies assumed, we cannot rule out the possibility that the choosiness of the constructions was recognized by some other means (Goldberg & Boyd, 2015; Yang, 2015). 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 new verb, and whether the item-specific behavior of this one verb is generalized to other new 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 Weak-Cx 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 Strong-Cx, 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.

3.1. 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.

3.2. 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 the Strong-Cx construction associated with a strong effect on the patient argument (strong-only verbs), and the other three verbs occurred exclusively in the Weak-Cx construction which was associated with a weak effect consistently for 2 of the 3 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 verb was witnessed in the "Weak-Cx" in both types of semantic contexts: contexts yielding a weak effect on the patient *and* con-

texts yielding a strong effect on the patient. In this way, the weak-effect semantics associated with the Weak-Cx was probabilistic, holding for all scenes associated with two verbs, and half of the scenes associated with a third verb. In order to facilitate detection of the distinctive behavior of the Weak-Cx-only verb by participants in the exposure phase, two instances of this verb were presented at the very beginning of the set, and another four sentences at the very end of the exposure set. This one verb was witnessed with a weak and a strong effect in succession, always in the same Weak-Cx (= word order PAV). The semantics associated with the Strong-Cx (= word order APV) was uniform: it was always associated with scenes that involved strong effects on the patient. The differences between the lexicalist condition of Experiment 1 and the preemption condition (Experiment 2) are summarized diagrammatically in Fig. 7.

3.3. Procedure

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

3.4. Results

3.4.1. 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.

Recall the input included 3 strong-only verbs, all witnessed having a strong effect on the patient argument; 2 weak-only verbs witnessed having a weak effect on the patient; and 1 (Weak-Cx-only) verb witnessed having both a strong and a weak effect on the patient. Two new novel verbs, for which no prior distributional information was provided in the input, were added in the test phase. The proportions of Weak-Cx and Strong-Cx word orders in subjects' productions are represented in Fig. 8. The labels Weak-Cx and Strong-Cx are used here and below to indicate which *word order* was produced (regardless of whether the effect on the patient was strong or weak).

If we consider the new novel verbs first (far right panel of Fig. 8), it is evident that the functions of the two constructions were readily detected, as they were in Experiment 1, since the novel verbs were used in whichever construction was better suited to the semantics of the scene being described, even though the weak-effect on the patient had only been probabilistically associated with the Weak-Cx during exposure.

Of special interest is that, unlike any verbs in Experiment 1, the Weak-Cx-only verb, which had been witnessed in both semantic contexts, shows a clear tendency to only be used in the Weak-Cx regardless of context (third panel from left). Thus, witnessing this verb in the Weak-Cx, even when the patient was strongly affected, appears to have statistically preempted this verb's use in the Strong-Cx. Moreover, participants tended to treat *all* verbs more lexically conservatively. That is, participants showed a tendency to use strong-only verbs with Strong-Cx word order, and weak-only verbs with Weak-Cx order, whether the effect on the patient argument in the scene being described was weak or strong. At the same time, the conservatism in the case of strong-only and weak-only verbs in Fig. 8 appears to be tempered by context, as is clear by comparing the right and left sides of each of the leftmost two panels; we return to this issue below (in Table 4).

[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.

The data was fit to a mixed effects logistic regression model similar to the one used in Experiment 1. The model predicts the occurrence of the Strong-Cx (=APV) order from the fixed predictors Effect and VerbType, with random intercepts for Subject, VerbForm, and VerbMeaning.[10] As in Experiment 1, sum contrast was used for all variables. We performed the same model selection procedure as in Experiment 1. Since no significant interaction between Effect and VerbType was found (see Appendix C), we removed the interaction term from the model. The fixed effects of the final model are reported in Table 3; likelihood ratio tests for the full model can be found in Appendix C.

We find a significant positive main effect of Effect, showing that participants tended to use the Strong-Cx (=APV) order when the patient was strongly affected. However, this tendency is balanced by significant effects of lexical conservativeness for each verb type: strong-only verbs tended to be used in the Strong-Cx word order (as shown by the positive estimate), and the weak-only verbs (including the Weak-Cx-only verb) tended to be used in the Weak-Cx (=PAV) word order (as shown by the negative estimate). As expected, new novel verbs are not produced in the Weak-Cx construction significantly more often than is found in the central tendency (when all verb types are combined). In a separate but similar mixed effects logistic regression restricted to the weak-only verbs and the preempted verb, no significant effect of VerbType was found ($\beta = 0.0652$, $SE = 0.2867$, $F = 0.227$, $p = 0.8202$), showing that the degree of lexical conservativeness was not measurably different between the two types of verbs witnessed only in the Weak-Cx word order.

In sum, it appears that participants were fairly lexically conservative with all verbs that were previously witnessed in the input. This is in stark contrast with Experiment 1, in which the general tendency was rather towards generalization according to constructional meaning, in both the lexicalist and alternating conditions. It thus seems that the presence of a preempted verb in the input, witnessed with the same construction in all contexts, encourages learners toward lexical conservativism across the board. At the same time, while participants 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 Strong-Cx word order—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 Weak-Cx word order—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, in the absence of verb-specific input, participants strongly tended to appropriately produce the Strong-Cx when there was a strong effect on the patient and the Weak-Cx when the effect on the patient was weak.

3.4.2. Comparing weak-only and strong-only verbs in the two experiments

In Experiments 1 and 2, the input for a subset of verbs was identical: strong-only verbs were only witnessed in the Strong-Cx with scenes in which the effect on the patient was strong, and weak-only verbs were only witnessed in the Weak-Cx with scenes in which the effect on the patient was weak. We can compare performance on just these verbs across the two experiments to determine whether participants were in fact more lexically

[10] In contrast with Experiment 1, the final model for Experiment 2 did converge even when the random intercepts for VerbForm and VerbMeaning were included, so they were kept in the model reported in Table 3. The standard deviations of the random effects were as follows: $SD_{\text{Subject}} = 0.9238$, $SD_{\text{VerbForm}} = 0.0002$, $SD_{\text{VerbMeaning}} = 0.1261$.

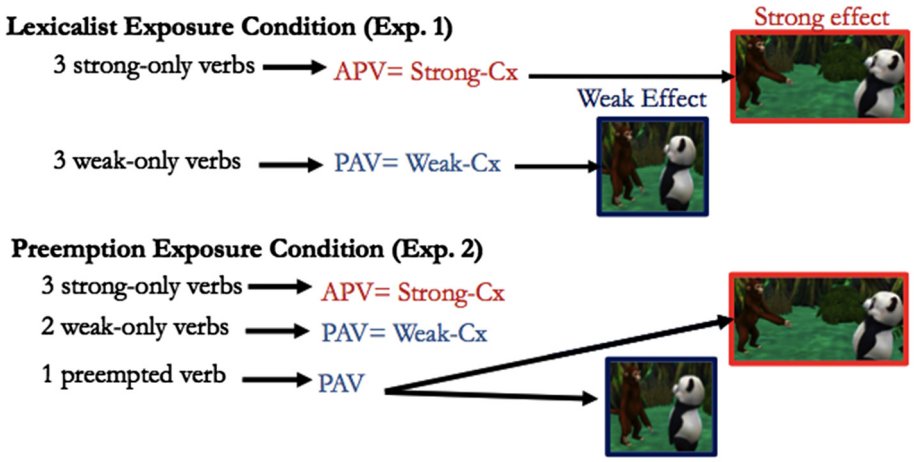


Fig. 7. Comparison of the types of exposure provided in the Lexicalist condition of Experiment 1 and the Preemption condition (Experiment 2). Exposure in Lexicalist and Preemption conditions was identical syntactically; only in the Preemption condition was one of the six verbs witnessed consistently in the PAV = "Weak-Cx" word order, even when the effect on the patient argument was strong.

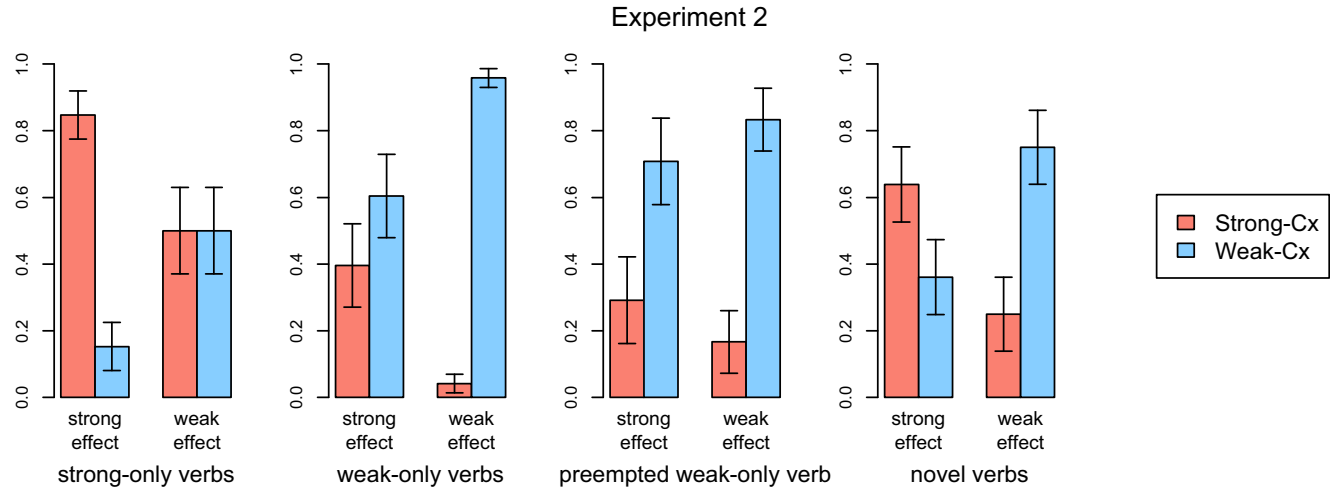


Fig. 8. Experiment 2: Proportions of Strong-Cx and Weak-Cx productions. Each of the four panels includes the proportion of participants' productions that describe scenes in which there was a strong effect on the patient (left side) or a weak effect on the patient (right side). Proportions of Strong-Cx (=APV) and Weak-Cx (=PAV) produced for the three verbs that had been presented only in the Strong-Cx construction, the two verbs witnessed only in the Weak-Cx construction with a weak effect on the patient, the one preempted verb (which was witnessed in the Weak-Cx construction for both strong and weak effects), and the two new novel verbs.

Table 3
Experiment 2 production task: Fixed effects of the logistic regression model predicting the occurrence of the Strong-Cx. Classification accuracy = 76.34%, *AUC* = 85.78%. Marginal R^2 = 36.86%, Conditional R^2 = 50.05%. Model formula: StrongCx ~ Effect + VerbType + (1 | Subject) + (1 | VerbForm) + (1 | VerbMeaning).

	Estimate	Std. error	z-value	p-value
(Intercept)	−0.6931*	0.3123	−2.219	0.0265
Effect (strong)	1.0216***	0.1464	6.977	<0.0001
VerbType (strong-only)	1.7149***	0.2485	6.901	<0.0001
VerbType (weak-only)	−0.9436***	0.2595	−3.636	0.0003
VerbType (preempted, Weak-Cx-only)	−1.1136***	0.3369	−3.306	0.0009
VerbType (novel)	0.3422	0.2397	1.428	0.1534

The stars next to the estimates reflect the significance threshold of the p-value: *** for p < 0.001, ** for p < 0.01, * for p < 0.05.

conservative in Experiment 2 without including the preempted verb, and whether there was an influence of constructions' semantics (whether the effect was strong or weak) in Experiment 2 beyond that displayed by the novel verbs. Thus we submitted the data from the strong-only verbs and the weak-only verbs in the lexicalist condition of Experiment 1 and in the only condition of Experiment 2 (hereafter called the preemption condition) to mixed effects logistic regression modeling, as in Experiment 1. As in Experiment 1, all interactions were initially tested (see Appendix D), and only the significant ones were kept in the final model, the fixed effects of which are presented in Table 4. Sum contrast was used for all variables, and since all variables are binary, only

Table 4
Lexicalist condition of Experiment 1 compared with Experiment 2 ("preemption" condition). Fixed effects of the logistic regression model predicting the occurrence of the Strong-Cx construction with strong-only verbs and weak-only verbs. Classification accuracy = 77.33%, *AUC* = 83.89%. Marginal R^2 = 38.44%, Conditional R^2 = 42.20%. Model formula: StrongCx ~ Effect + VerbType * Condition + (1 | Subject) + (1 | VerbForm) + (1 | VerbMeaning).

	Estimate	Std. error	z-value	p-value
(Intercept)	−0.2627	0.1646	−1.596	0.1105
Condition (preemption)	−0.0330	0.1375	−0.240	0.8104
Effect (strong)	1.1561 ***	0.1205	9.591	<0.0001
VerbType (strong-only)	0.8779 ***	0.1241	7.074	<0.0001
Condition (preemption) × VerbType (strong-only)	0.4402 ***	0.1160	3.796	0.0001

The stars next to the estimates reflect the significance threshold of the p-value: *** for p < 0.001, ** for p < 0.01, * for p < 0.05.

one level of each variable is reported. As previously, random intercepts for subjects, verb forms, and verb meanings were included.[11]

There are again significant main effects of Effect and VerbType, and only the latter is involved in a significant interaction with Condition. The interaction means that the effect of VerbType is significantly stronger in the preemption condition than in the lexicalist condition, which confirms that participants in the preemption condition displayed a stronger tendency to use strong-only verbs and weak-only verbs in the Strong-Cx and Weak-Cx constructions, respectively. Thus, witnessing a single verb used in one word order, regardless of whether the effect on the patient argument was strong or weak, reduced the tendency to use the other verbs on the basis of the constructions' semantics. This is consistent with Wonnacott et al.,'s (2008) observation that learners decide whether verbs can be used flexibly partly on the basis of whether other verbs have been witnessed being used flexibly. Here we find evidence that all verbs are more lexically conservative, when there is evidence that a single verb loyally appears in the Weak-Cx word order irrespective of whether the effect on the patient was strong or weak. The results of Experiment 2 are also consistent with an interpretation in terms of cue validity, since participants were more lexically conservative when verbs predicted the choice of construction perfectly (probability = 1) and the effect on the patient only predicted the choice of construction with a probability of 0.92.

At the same time, the significant influence of Effect (whether strong or weak) does not interact with Condition, which suggests that participants in both experiments were influenced by the type of scene being described in selecting which construction to produce. In the case of Experiment 2, this suggests that both factors played a role: the verbs that had only been witnessed in the Strong-Cx or only in the Weak-Cx showed a strong tendency to be used in their respective constructions, but participants were also somewhat influenced in their choice of construction by whether the scenes at test involved a strong or weak effect on the patient argument.

3.4.3. 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 Fig. 9 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.

Judgment results are consistent with the production results just reviewed. All three types of verbs were judged to be more acceptable when they were used the way they had been witnessed, regardless of the effect on the patient; i.e., strong-only verbs tended to be judged acceptable in the Strong-Cx and unacceptable in the Weak-Cx, and vice versa for all weak-only verbs and for the preempted verb. 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 Strong-Cx and weak effect for the Weak-Cx), and lower when the construction was used with an incongruent scene.

As with the production data, we pooled the sentence rating data from Experiment 2 with that of the lexicalist condition of Experiment 1, in order to test whether the presence of a preempted verb has a significant impact on the factors that influence sentence acceptability. To test for significance, we submitted the data to mixed effects linear regression. The dependent variable in the model is the z-score sentence rating submitted for each trial, and the main predictors are Condition (lexicalist vs. preemption), EffectCongruent and VerbCongruent. The latter two are binary variables that respectively indicate whether the trial sentence uses the construction that is congruent with the effect on the patient as displayed in the accompanying video (Strong-Cx [=APV word order] for strong effect and Weak-Cx [=PAV order] for weak effect), and whether the verb is used in the construction it was consistently witnessed with in the input (Strong-Cx for strong-only verbs and Weak-Cx for both weak-only verbs and the preempted Weak-Cx-only verb). Sum contrast was used for all variables. Subject, VerbForm, and VerbMeaning were again included as random factors, but they did not capture significant variance (all < 0.0001). Two-way interactions between all factors were also included; since no significant interaction was found between EffectCongruent and VerbCongruent, it was removed from the final model. The fixed effects of the final model are reported in Table 5; the full model can be found in Appendix E.

Both types of congruency have a significant main effect on sentence ratings, showing that both factors contribute to sentence acceptability in both conditions. However, both are also involved in significant interactions with Condition: in the preemption condition (Experiment 2), the effect of EffectCongruent is markedly weaker, while that of VerbCongruent is markedly stronger. 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.

In sum, the results of the sentence rating task are consistent 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 "Weak"-Cx (=PAV order) regardless of whether the effect on the patient was strong or weak.

4. General discussion

The results of Experiment 1 confirm the idea that participants readily learn the functions of individual constructions, and readily generalize a construction for use with new verbs, if the function of

[11] The standard deviations of the random effects were as follows: $SD_{Subject}$ = 0.3897, $SD_{VerbForm}$ = 0.1853, $SD_{VerbMeaning}$ = 0.1663.

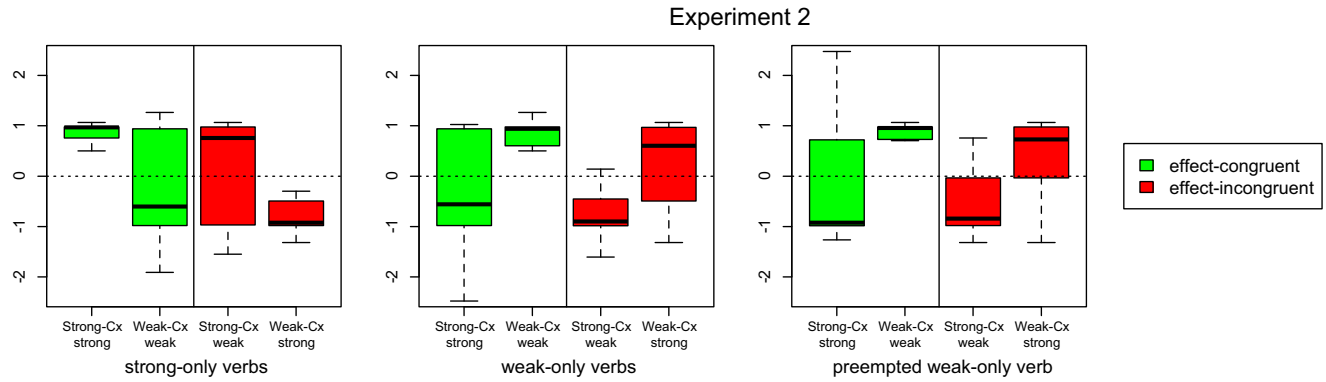


Fig. 9. Experiment 2. Box plots of the distribution of grammaticality ratings (z-scores) for each verb type, and each combination of construction and effect on the patient. Combinations that were congruent with the input as regards the effect on the patient (i.e., Strong-Cx with strong effect, Weak-Cx with weak effect) are plotted on the left-hand side of each box (in green); the other, incongruent combinations are plotted on the right-hand side (in red). Outliers are not plotted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Fixed effects of the linear regression model predicting the z-score ratings provided by subjects in the sentence rating task of Experiment 2 (preemption condition), compared to the lexicalist condition of Experiment 1. Marginal R^2 = 22.59%, Conditional R^2 = 22.59%. Model formula: Zscore ~ Condition * EffectCongruent + Condition * VerbCongruent + (1 | Subject) + (1 | VerbForm) + (1 | VerbMeaning).

	Estimate	Std. error	t-value	p-value
(Intercept)	<0.0001	0.0360	0.000	1
Condition (preemption)	0.0000	0.0360	0.000	1
EffectCongruent (true)	0.3033***	0.0360	8.416	<0.0001
VerbCongruent (true)	0.3211***	0.0360	8.909	<0.0001
Condition (preemption) × EffectCongruent (true)	−0.1104**	0.0360	−3.062	0.0023
Condition (preemption) × VerbCongruent (true)	0.1031**	0.0360	2.861	0.0044

The stars next to the estimates reflect the significance threshold of the p-value: *** for p < 0.001, ** for p < 0.01, * for p < 0.05.

that construction is better suited to convey the intended message. When both verbs and the type of event (strong or weak effect on the patient) 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 is 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 patient 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 patient. In the alternating condition of Experiment 1, two out of six of the verbs were unreliable predictors of which construction would be used, as both were used equally often in the two constructions (with corresponding differences in the degree of effect). In this case, participants *entirely* ignored the distribution of all six verbs in the input, using all of the verbs freely in either construction, depending only on whether the effect on the patient was strong or weak.

Results from Experiment 2 add important nuance to the finding that participants readily generalize beyond their input. When learners witnessed a verb being used in one construction to describe *either type of message*, they restricted that verb to that construction for either type of message. The verb is 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, and the influence of constructional meaning was in evidence as well, particularly in 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 in Experiment 1, even though they largely ignored them. 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 tend to use all verbs in whichever construction they had been witnessed, even though they demonstrated an appreciation of the meanings of the constructions, particularly with new novel verbs. The tendency to infer from evidence that a single verb is statistically preempted from occurring in a construction, that other verbs are also restricted, demonstrates the power of statistical preemption, particularly since learners had no prior knowledge that would lead them to expect verbs to be lexically restricted.

At the same time, it could be that learners would have displayed, in Experiment 2, a greater willingness to generalize verbs that had not specifically been preempted from appearing in the alternative, if the communicative stakes had been higher. In both of the present experiments, participants were only asked to describe each scene; there was no communicative pressure to convey whether the effect on the patient argument was strong or weak, and no reward for doing so. Given the task demands, using each verb in the same construction it had been witnessed in was a safe response. If, instead, communicative success demanded that

the strength of effect was communicated, it is quite possible that learners would have shown an even greater propensity to take advantage of the semantic distinction between the two constructions. It is notable in this respect that participants elected to generalize so readily in Experiment 1, disregarding verb-specific input and taking full advantage of the functional distinction between the two constructions to convey a semantic distinction.

Additional work is needed to investigate whether children are as sensitive to the semantics/functions associated with abstract constructions as adults, and whether they are as sensitive to statistical preemption as adults. With sufficient input (Wonnacott, Boyd, Thompson, & Goldberg, 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 older children and adults are (e.g., Akhtar, 1999; Boyd, Gottschalk, & Goldberg, 2009; Theakston, 2004; Tomasello, 2000; Tomasello, 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 good deal more input than adults before they take advantage of the sort of indirect negative evidence that statistical preemption provides (Brooks & Tomasello, 1999; Goldberg & Boyd, 2015; Hao, 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 in the first experiment, as speakers selected whichever of the two phrasal constructions better suited their intended message. In the lexicalist condition, in which verbs and semantic scenes were both perfect predictors of the choice of construction used, 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.

Of particular interest in Experiment 2 is the finding that witnessing a single verb (out of six) stubbornly occurring in one construction, even when the semantics of the scene better matched the other construction, was found to increase participants' tendency to restrict the distribution of all verbs. Participants behaved markedly more conservatively than in either condition of Experiment 1. At the same time, participants in Experiment 2 also demonstrated a sensitivity to the semantics of the learned constructions, particularly in their use of novel verbs. 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.

We also find evidence that is consistent with a role for cue validity in determining whether constructions are used productively (Thothathiri & Rattinger, 2016). There was a significant difference between the lexicalist and alternating conditions in Experiment 1; 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. In this case, verbs were less reliable cues to which construction should be used than were the type of scenes. We also saw an effect that is consistent with cue validity 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. At the same time, we have seen that participants are not using cue validities in a blind fashion, without regard to expressive power, since the cue validities in the lexicalist condition of Experiment 1 controlled for the cue validities of verbs and scenes, and yet the effect of scene was stronger than the effect of verb in participants' choice of construction (see also Perek & Goldberg, 2015).

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 reduced 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 (cf. Boyd & Goldberg, 2011; Brooks & Tomasello, 1999).

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 wish to thank three anonymous reviewers and Mike Frank for very helpful feedback on a previous version of this paper. We are also grateful to Isaac Treves for running participants. We thank Nicholas Groom, Rachel Hatchard, Jeannette Littlemore, and Bodo Winter, for their comments on an earlier version of this paper, and to Sammy Floyd, Karina Tachihara, and Libby Barak for useful discussion. This paper also greatly benefitted from statistical advice from Bodo Winter. We are responsible for any remaining errors.

Appendix A

Likelihood ratio tests for the full mixed effects regression model corresponding to the final model reported in Table 1:

Effect	df	Chi-square	p-value
Condition	1	7.50**	0.006
Effect	1	239.42***	<0.0001
VerbType	2	15.67***	0.0004
Condition × Effect	1	7.39**	0.007
Condition × VerbType	2	0.15	0.93
Effect × VerbType	2	0.82	0.66
Condition × Effect × VerbType	2	2.13	0.35

Appendix B

Likelihood ratio tests for the full mixed effects regression model corresponding to the final model reported in Table 2:

Effect	df	Chi-square	p-value
EffectCongruent	1	149.73***	<0.0001
Condition	1	0.04	0.85
VerbCongruent	1	12.52***	0.0004
EffectCongruent × Condition	1	3.26	0.07
EffectCongruent × VerbCongruent	1	0.21	0.65
Condition × VerbCongruent	1	5.73*	0.02
EffectCongruent × Condition × VerbCongruent	1	0.55	0.46

Appendix C

Likelihood ratio tests for the full mixed effects regression model corresponding to the final model reported in Table 3:

Effect	df	Chi-square	p-value
Effect	1	44.25***	<0.0001
VerbType	3	77.13***	<0.0001
Effect × VerbType	3	2.81	0.42

Appendix D

Likelihood ratio tests for the full mixed effects regression model corresponding to the final model reported in Table 4:

Effect	df	Chi-square	p-value
Condition	1	0.25	0.62
VerbType	1	59.21***	<0.0001
Effect	1	111.04***	<0.0001
Condition × VerbType	1	11.65***	0.0006
Condition × Effect	1	0.18	0.67
VerbType × Effect	1	1.99	0.16
Condition × VerbType × Effect	1	0.3	0.58

Appendix E

Likelihood ratio tests for the full mixed effects regression model corresponding to the final model reported in Table 5:

Effect	df	Chi-square	p-value
Condition	1	0	>0.99
EffectCongruent	1	67.47***	<0.0001
VerbCongruent	1	75.10***	<0.0001
Condition × EffectCongruent	1	9.40**	0.002
Condition × VerbCongruent	1	8.21**	0.004
EffectCongruent × VerbCongruent	1	0.04	0.85
Condition × EffectCongruent × VerbCongruent	1	0.01	0.92

Appendix F. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2017.06.019.

References

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language, 26*(2), 339–356.

Amato, M. S., & MacDonald, M. C. (2010). Sentence processing in an artificial language: Learning and using combinatorial constraints. *Cognition, 116*(1), 143–148. http://dx.doi.org/10.1016/j.cognition.2010.04.001.

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., Freudenthal, D., & Chang, F. (2014). Avoiding dative overgeneralisation errors: Semantics, statistics or both? *Language, Cognition and Neuroscience, 29*(2), 218–243.

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baker, C. L. (1970). Notes on the description of English questions: The role of an abstract question morpheme. *Foundations of language*, 197–219.

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

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=lme4ii.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. *The Crosslinguistic Study of Sentence Processing, 3*, 73–112.

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.

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

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.

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.

Bowerman, M. (2000). Where do children's word meanings come from? Rethinking the role of cognition in early semantic development. In L. Nucci, G. Saxe, & E. Turiel (Eds.), *Culture, thought and development* (pp. 199–230). Mahwah, NJ: Lawrence Erlbaum.

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. (1990). The "natural approach" to reasoning. In W. Overton (Ed.), *Reasoning, necessity and logic*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

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.

Bresnan, J. (2011). Linguistic uncertainty and the knowledge of knowledge. In Roger Porter & Robert Reynolds (Eds.), *Thinking Reed* (pp. 69–75). Reed College, Portland, OR: Centenial Essays by Graduates of Reed College.

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

Brooks, P. J., Braine, M. D., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language, 32*, 76–95.

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping of a phrasal form and meaning. *Developmental Science, 8*, 500–508.

Chan, A., Lieven, E., & Tomasello, M. (2009). Children's understanding of the agent-patient relations in the transitive construction: Cross-linguistic comparisons between Cantonese, German, and English. *Cognitive Linguistics, 20*(2), 267–300.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

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

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*.

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.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science, 17*, 684–691.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

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., & Boyd, J. K. (2015). A-adjectives, statistical preemption, and the evidence: Reply to Yang (2015). *Language, 91*(4), e184–e197.

Goldberg, A. E., Casenhiser, D., & Sethuraman, N. (2005). The role of prediction in construction learning. *Journal of Child Language, 32*(2), 407–426.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*(5), 178–186.

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

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

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

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

MacWhinney, B. (2012). The logic of the unified model. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 211–227). London & New York: Routledge.

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.

Naigles, L. (2000). Manipulating the input: Studies in mental verb acquisition. In B. Landau & Sabini et al. (Eds), Perception, cognition, and language: Essays in honor of Henry and Lila Gleitman (pp. 245–274). Cambridge, MA: MIT Press.

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.

Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam: Benjamins.

Perek, F., & Goldberg, A. E. (2015). Generalizing beyond the input: The functions of the constructions matter. *Journal of Memory and Language, 84*, 108–127.

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.

Robenalt, C., & Goldberg, A. E. (2015). Judgment 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, 26*(3), 467–503.

SAS Institute Inc (1978). *SAS Technical Report R-101. Tests of Hypotheses in Fixed-Effects Linear Models*. Cary, NC: SAS Institute Inc.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition, 113*(2), 226–233.

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.

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

Tomasello, Michael. (2003). *Constructing a language*. Harvard University Press.

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

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., Newport, E., & Tanenhaus, M. (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.