Generalizing beyond the input

Perek, Florent; Goldberg, Adele

DOI:
10.1016/j.jml.2015.04.006

Document Version
Peer reviewed version

Citation for published version (Harvard):
Perek, F & Goldberg, A 2015, 'Generalizing beyond the input: The functions of the constructions matter', *Journal of Memory and Language*, vol. 84, pp. 108-127. https://doi.org/10.1016/j.jml.2015.04.006

Generalizing beyond the input: the functions of the constructions matter

Florent Perek & Adele E. Goldberg

A growing emphasis on statistics in language learning raises the question of whether learning a language consists wholly in extracting statistical regularities from the input. In this paper we explore the hypothesis that the functions of learned constructions can lead learners to use language in ways that go beyond the statistical regularities that have been witnessed. The present work exposes adults to two novel word order constructions that differed in terms of their functions: one construction but not the other was exclusively used with pronoun undergoers. In Experiment 1, participants in a lexicalist condition witnessed three novel verbs used exclusively in one construction and three exclusively in the other construction; a distinct group, the alternating condition, witnessed two verbs occurring in both constructions and two other verbs in each of the constructions exclusively. Production and judgment results demonstrate that participants in the alternating condition accepted all verbs in whichever construction was more appropriate, even though they had seen just two out of six verbs alternating. The lexicalist group was somewhat less productive, but even they displayed a tendency to extend verbs to new uses. Thus participants tended to generalize the constructions for use in appropriate discourse contexts, ignoring evidence of verb-specific behavior, especially when even a minority of verbs were witnessed alternating. A second experiment demonstrated that participants' behavior was not likely due to an inability to learn which verbs had occurred in which constructions. Our results suggest that construction learning involves an interaction of witnessed usage together with the functions of the constructions involved.

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

1. Introduction

There is a growing body of research demonstrating that children and adults are acutely sensitive to the statistical properties of the language that they witness, insofar as a great deal of "item-specific" statistical information about particular words is recorded. In fact, the token frequencies of words and phrases play a key role in a number of linguistic

processes (e.g., Bybee 2010; Ellis 2002; Gibson, Schutze, & Salomon 1996; Gries & Divjak 2012). For example, frequent subject auxiliary combinations are more likely to be produced earlier by children than less frequent combinations (Dąbrowska & Lieven 2005). Sentences tend to be comprehended more quickly when individual verbs appear with complements that are statistically more likely (Ford et al. 1982; Garnsey et al. 1997; MacDonald et al. 1994). More frequent combinations of words are more likely to be reduced and/or grammaticalized (Bybee & Hopper 2001; Gahl & Garnsey 2004; Kuperman & Bresnan 2012), are processed faster (Arnon and Snider 2010; Gathercole & Baddeley 1993), and are repeated faster and more accurately, both by children and adults (Bannard & Matthews 2008; Bod 1998).

Statistical distributional information is recognized as a rich source of evidence available to human and machine learners alike. Transitional probabilities, known to be tracked even by young infants (Saffran, Aslin, and Newport 1996), can be used to learn higher order generalizations akin to phrase structure rules (Saffran 2001; 2003). With the wide availability of large corpora, researchers are using co-occurrence statistics as a way of discovering semantic structure as well as formal regularities (Lund and Burgess 1996; Landauer et al. 1998; Baroni & Lenci 2010; Mintz et al. 2002). It has been demonstrated that a combination of type and token frequencies in the input plays an important role in whether learners generalize a novel word or novel grammatical construction beyond their exposure, productively applying it to new instances (e.g., Barðdal 2008; Casenhiser & Goldberg 2005; Suttle & Goldberg 2011; Wonnacott, Newport & Tanenhaus 2008; Wonnacott, Boyd, Thompson & Goldberg 2012; Xu & Tanenbaum 2007).

In an important study that inspired the present work, Wonnacott et al. (2008) found that the overall statistics of an artificial language plays a role in how individual items are treated. Artificial language learning experiments involve having participants learn a miniature language by exposing them to a set of novel phrases or sentences that are paired with some sort of interpretation (e.g., Goldberg, Casenhiser, & Sethuraman 2004; Casenhiser & Goldberg 2005; Culbertson, Legendre, & Smolensky 2012; Wonnacott, Newport & Tanenhaus 2008). Wonnacott et al. demonstrated that adult learners exposed to a "lexicalist" language in which most verbs appeared in only one construction behaved conservatively, avoiding extending verbs for use in a different construction; on the other hand, learners exposed to a "generalist" language in which the majority of verbs alternated, appearing in both of two constructions, readily assumed that all verbs alternated. Wonnacott (2011) is a similar study that has replicated the basic findings with children.

These sorts of input-driven findings may seem to lead to the conclusion that learners acquire their knowledge of language from simply gleaning statistical regularities from the language input. Is it possible that language learning is *wholly* a process of learning various statistical regularities in the input (e.g., Taylor 2012)? Weighing against this conclusion is a range of findings that indicate that learners bring to the task of language learning certain biases that help shape what is learned. While certain domain-specific "substantive" biases have been proposed (e.g., Culbertson et al. 2012; but see Goldberg 2013), other biases have been argued to emerge from the communicative function of language (Levy & Jaeger 2007; Jaeger 2010; Hawkins 1994, 2004, 2014; Mahowald, Fedorenko, Piantadosi, & Gibson 2013; Piantadosi, Tily, & Gibson 2012), from domain general constraints on working memory (Gathercole & Baddeley 1993;

Hudson Kam & Newport 2005; Fedorenko, Gibson & Rohde 2006), for a preference for simplicity (Culbertson & Newport 2015), or from rational inductive processes (Griffiths et al. 2010; Perfors, Tenenbaum, & Wonnacott 2010). It is also well-established that the meanings of words play a role in constraining their distributions and *vice versa*, insofar as semantically related words tend to occur in similar distributional contexts (Arunachalam, & Waxman 2014; Fisher, Gleitman & Gleitman 1991; Scott & Fisher 2009; Waxman et al. 2009). Somewhat less emphasized have been constraints that emerge from the function of particular constructions (but see e.g., Ambridge et al. 2008; Ambridge & Goldberg 2008; Bybee 1985; Lakoff 1987; Langacker 1987). This is the focus of the present work.

We begin by taking a closer look at the Wonnacott et al. (2008) experiments that had shown that the statistical properties of an artificial language as a whole determined how individual words were used. Participants were taught five novel nouns and 12 novel verbs and were then exposed, over a five-day period, to a language that contained two constructions with different word orders, VSO (Verb Subject Object) and VOS-ka (Verb Object Subject followed by a particle, *ka*). In two experiments, the proportion of verbs that occurred in either construction was varied. Relevantly to our purposes, in their first experiment, a third of the verbs (4) occurred only in the VSO construction, a third of the verbs occurred only in the VOS-ka construction, and the final third appeared in both constructions with equal probability; in this case, participants tended to be lexically conservative, preferring to produce and expecting to hear the one-construction verbs in the construction that they had witnessed those verbs in.[1] The second experiment included twice the number of alternating verbs (two thirds: 8 verbs) as non-alternating verbs (2 verbs were witnessed only in VSO and 2 only in VOS-ka); in this case, there was a much stronger tendency to use all of the verbs in both constructions. Thus learners made use of not only the behavior of individual verbs, but also more general patterns in the input. They appear to implicitly assume roughly, "if few verbs alternate, I will be conservative and only use verbs as I have witnessed them; but if most verbs alternate, perhaps all of the verbs alternate."

A final experiment compared contrasting inputs. One group witnessed a completely alternating language in that all 8 verbs alternated (in a ratio of 7-1 in favor of the VOS-ka construction over the VSO construction). The other group witnessed a completely lexicalist language: 7 verbs appeared only in the VOS-ka construction and 1 verb appeared only in the VSO construction. Subjects learning the alternating language alternated at roughly the same 7-1 rate, even for novel verbs. The learners of the lexicalist language were lexically conservative, using the VOS-ka verbs in that pattern and the VSO verb in its pattern. In other words, in the absence of alternating verbs in the input, learners tended to assume that no verbs could alternate.

It is important to note that the two constructions used in the Wonnacott et al. (2008) experiments were interchangeable, in that there was no discernible difference in their meanings or discourse functions. This situation rarely occurs in natural languages; whenever there exist verbs that alternate between two constructions, there is almost

[1]Wonnacott et al. (2008) also varied the token frequency of different verbs, and found entrenchment effects: the more frequent verbs were less likely to be used in a construction they had not been witnessed in, and were judged to be less acceptable. As the focus of the present work was on the functions of constructions, we did not include a token frequency manipulation.

always a functional difference between the constructions. If the constructions do not differ in terms of truth conditions, then they involve a distinction in terms of construal, information structure, pragmatics, register, or dialect (e.g., Bolinger 1968; Goldberg 2004; Langacker 1987). For example, the English double-object construction (e.g., *She gave him a book*) and *to*-dative (e.g., *She gave a book to him*) are a classic case of an alternation, in that many verbs can appear in either construction, and as expected, there are well-established information structure differences between the two (Thompson 1990; Green 1974; Oehrle 1976; Goldberg 1995; Rappaport Hovav & Levin 2008; Bresnan & Ford 2010). In particular, the double-object construction is more likely to be used when the recipient (here, *him*) is more given and topical in the discourse than the theme (here, *a book*); the *to*-dative construction is much less constrained, and, moreover, can be used to convey caused-motion in addition to transfer of possession.

The present study differs from Wonnacott et al. (2008) in that it aims to present learners with a more ecologically valid situation, in which novel verbs are used in two constructions that differ in terms of word order *and* information structure properties. In particular, one construction is always used with a pronominal undergoer argument ($\text{Pronoun}_{\text{Undergoer}}$ NP_{Agent} V: a ProSV construction), while the other occurs exclusively with lexical noun phrase arguments in a differing order (NP_{Agent} $\text{NP}_{\text{Undergoer}}$ V: an SOV construction). Our main goal is to determine how learners weigh the function of a construction in the face of potentially conflicting distributional properties of the verbs in the input. That is, if speakers associate a new construction with a particular function, do they then use it whenever it is appropriate and essentially disregard verb-specific distributional information? Or do learners take their cues about usage entirely from verbal distribution, failing to use a new construction in an appropriate context unless the input has licensed the generalization?

2. Experiment 1

In Experiment 1 we compare participants' production and judgment data in two groups. In a lexicalist condition, half of the novel verbs are witnessed in one construction, and half in a distinct construction, throughout the exposure phase. The two constructions differ in terms of word order and in whether pronouns are used. In an alternating condition, one third of the novel verbs are used in both constructions during exposure, another third is witnessed in only one construction, and the final third is only witnessed in the other construction. At test, different questions are asked to elicit productions, such that the questions biased the discourse context in that one or the other construction was better suited.

If speakers base their productions solely on the basis of distributional evidence in the input, we might expect that if the input suggests that a verb occurs in only one construction, speakers should restrict their productions to use that verb only in that construction. If, however, speakers prefer to use constructions that are better suited to the discourse, speakers may display a tendency to disregard verb-specific distributional evidence in the input. It is also possible that speakers are capable of using both factors, in which case we might see some degree of lexical conservatism—using verbs in the constructions in which they had been witnessed being used—and some degree of sensitivity to the function of constructions. The relative weight of the two factors may

also be influenced by more general facts about the language such as whether any verbs are witnessed appearing in both of the available constructions.

2.1 Participants

Participants were 42 undergraduate students in the department of psychology at Princeton University who participated in the experiment for course credit. All were native speakers of English and had normal or corrected vision. Twenty-four (17 female, 7 male, aged 18-22, mean 19.5) were tested in the lexicalist condition further divided into two groups as described below, and 18 (12 female, 6 male, aged 18-22, mean 19.8) were tested in the alternating condition.

2.2 Materials

Word order in the artificial language departed from standard English syntax, and consisted of two constructions involving different orders: SOV (NP_{Agent} $NP_{Undergoer}$ V) and OSV ($Pronoun_{Undergoer}$ NP_{Agent} V). Crucially, the two constructions further differed in that the former was always used with both arguments expressed as lexical noun phrases, while the latter was always used with a lexical noun phrase agent and a pronominal undergoer; hence, the OSV construction is hereafter referred to as "ProSV."

These two constructions provided participants with two possible ways to describe a given transitive scene. For example, assuming the meaning "punch" is assigned to the verb form *moop*, the sentence meaning given in (1) below can be encoded with the SOV construction, as in (2), or with the ProSV construction as in (3), depending on whether the speaker refers to the undergoer with a pronoun:

(1) 'RABBIT$_{(agent)}$ PUNCH MONKEY$_{(Undergoer)}$'
(2) the rabbit the monkey mooped (SOV construction)
(3) him the rabbit mooped (ProSV construction)

The NP form of the arguments in the stimuli always consisted of a regular English definite noun phrase (i.e., the article *the* followed by a noun), while the pronominal form was always the pronoun *him*. Novel verbs were used in either the past tense (*-ed*) or progressive form (*-ing*), as in English.

Separate groups of participants were assigned to one of two exposure conditions. In the lexicalist condition, each verb was witnessed in a single construction throughout the exposure set; three verbs occurred exclusively in the SOV construction, and three in the ProSV construction. In the alternating condition, two of the verbs were witnessed in both constructions (50% of the time in each), two verbs occurred exclusively in the SOV construction, and two occurred exclusively in the ProSV construction. In other words, the two input conditions differed in terms of the presence or absence of any alternating verbs. The proportion of verbs that alternated even in the "alternating" condition was only one third. The assignment of novel verbs to each class was randomly determined for each participant.

Participants in the lexicalist condition were further divided into two roughly equal groups that were shown slightly different materials in the exposure set. For thirteen participants, one of two suffixes, *-o* and *-ee*, was added to each verb stem, such that all verbs occurring in the same construction had the same suffix. For the other 11

participants, no morphological indicators of the construction were used. We suspected that morphological marking might draw attention to the verbs and help participants to track their distributional properties; we thus expected more lexical conservatism from participants exposed to morphologically marked verbs. However, there turned out to be no significant differences correlated with verb marking in either the production or judgment task—neither in terms of lexical conservatism nor in terms of the effect of context—so we collapse the two groups into a single group in what follows.

The lexicon of the artificial language included six English names for animals (*cat*, *monkey*, *panda*, *pig*, *rabbit*, *wolf*) and eight novel 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 verbs were used as novel verbs in the test phase (cf. 2.3.1), in order to assess how learners treat items for which they did not receive any prior distributional information.

Each verb described one of eight different transitive actions enacted by anthropomorphized animals in 3D animations presented in video clips (cf. Section 2.3):[2] HEADBUTT, HUG, PULL, PUNCH, PUSH, SLAP (with both hands), SPIN (spin towards and hit the undergoer), SWIRL-STRIKE (strike with a swirling blow). The videos depicting each action presented the event as one in which the agent performs different gestures resulting in different effects on the undergoer (such as moving backwards, bending one's back, wobbling, jumping up and down, lifting one's arm and shaking one's head, etc.). The assignment of verb forms to verb meanings was randomized for each participant.

As described in more detail below, participants were then asked to produce sentences in discourse contexts that could bias them to treat the undergoer argument as either discourse-given, which would favor the ProSV construction; or discourse-new, which would favor the SOV construction. Hence, our experiment builds on participants' prior knowledge of when pronouns are appropriate, allowing participants to associate the novel constructions with distinct discourse functions. In particular, we rely on the fact that pronouns refer to discourse-given and topical arguments, and on the expectation that participants will transfer this knowledge to the artificial language. We also collected acceptability ratings.

2.3 Procedure

All the instructions were given in written form on the computer screen. For each participant, the experiment was conducted over two identical sessions on two consecutive days. Each session was divided in two parts: an exposure phase and a test phase. All artificial language sentences used in the experiment, as well as the questions asked during the test phase, were presented to subjects both in the visual and auditory modalities: sentences were shown on the screen and spoken by a computer-generated voice. The experiment was entirely implemented as a computer task with the PsychoPy software

[2] The computer animations were created with Alice (http://www.alice.org), a visual programming language designed for educational purposes. Originally intended for teaching object-oriented programming, Alice includes a development environment that allows users to easily create 3D-animated "virtual worlds" in which agents can be programmed to move and act in particular ways by means of a simple "point-and-click" interface. Although 3D animation is merely a means to an end for Alice's intended purpose, we took advantage of its powerful features that require little prior skills in order to design our visual stimuli.

package (Pierce 2007). The sound files for all sentences were generated using the speech synthesis features of Mac OS X 10.9 ("say" command).

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

In the last step of the exposure phase, the participants were shown three blocks of scenes matched with a sentence description, and were instructed to repeat each sentence out loud. The sentence description was shown on the screen and was also spoken by a computer-generated voice. Each verb was used twice in each block, each time with a different, randomly selected pair of characters. Each exposure set thus comprised 12 sentences per block, for a total of 36 input sentences.

The test phase, described in detail below, included a production task followed by a sentence-rating task.

2.3.1 Production task

The production task contained 16 triples consisting of 1) a vocabulary question, 2) a sentence comprehension question and 3) a sentence production question (always in that order). Our key dependent measure was the production data. The point of including and interspersing the other tasks was twofold. First, the tasks act as distractors for the production task, which should dampen the influence of production-to-production priming (i.e., using the same structure over and over in a number of consecutive trials). Second, the comprehension task in particular presents subjects with additional sentences, which should remind them of the range of constructions that the language contains and further serve to counter production-to-production priming.

The participants were told to imagine that they had to teach the language to another person by providing information as requested. To convey the fact that the person asking the questions was distinct from the person providing the input (the latter knowing the language, the former not), we used a different computer voice for the exposure phase and the test phase (a male voice, "Tom", and a female voice, "Samantha," respectively).

Sentence production task (question manipulation): This task provided our key dependent measure. Participants were asked to describe a scene displayed on the screen, using the artificial language. To facilitate the production task, the verb was provided (in the past tense form) in written form on the computer screen.

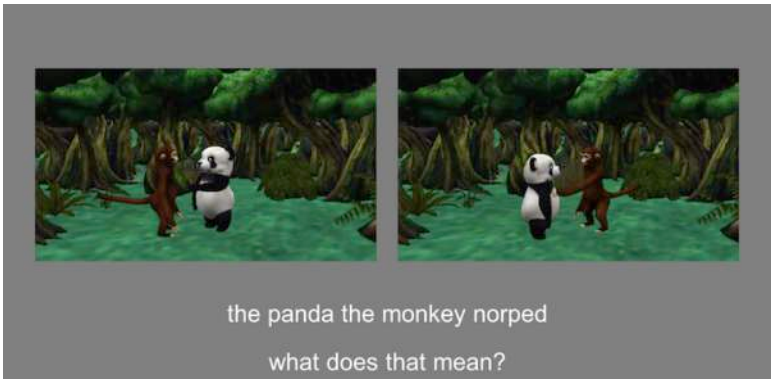
Questions were systematically varied in order to elicit a response that included a pronoun or a lexical NP. The NP-biased context question did not mention either participant: i.e., "what happened here?" The pronoun-biased context question mentioned the undergoer participant: i.e., "what happened to the rabbit/monkey/etc.?" In the former case, both characters are discourse-new and should therefore be referred to in their full NP form; in the latter case, the fact that the undergoer is mentioned in the question makes it the topic of the current discourse; therefore a contextually appropriate response encourages speakers to refer to it with a pronoun. In addition, the scenes intended to provide an NP-biased context contained two additional animals that were not doing anything but whose presence was meant to make a pronoun referent potentially ambiguous, in order to further promote the use of lexical NPs for both arguments.

Importantly, the two kinds of presentation context do not constitute inviolable constraints on the selection of referential forms. While it is not ideal from a pragmatic perspective to use a noun phrase to refer to a character that has just been mentioned, or, conversely, to use a pronoun to refer to a character that has not been recently mentioned, using the other form is not ungrammatical, and speakers did sometimes deviate from the pragmatic norm. In fact, they did so at times due to verb-specific factors, as discussed below. It is also relevant that speakers could have potentially produced pronouns using the SOV order (producing SProV utterances) instead of using only the SOV and ProSV constructions that had been instantiated in the input.

All six verbs introduced during the exposure phase, as well as two new novel verbs, were presented twice during the production task, once in a NP-biased context and once in a pronoun-biased context, each time with a different pair of agent and undergoer arguments which had not been used with this verb during the exposure phase. In all tasks, the left-to-right orientation of the undergoer and agent in the scene was randomly determined for each trial, with the agent presented on the right in half the scenes and on the left in the other half. The participants' responses to the production task in each trial were recorded into separate sound files using a regular laptop microphone.

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

Vocabulary distractor task



Comprehension distractor task

Production task

Figure 1: Screenshots of the three tasks given in each comprehension/production test triple. Testing consisted of 16 such triples.

Vocabulary distractor task: Participants were asked to identify the correct label for a given action shown on the screen (i.e., a verb), when given two alternatives. The question was of the form "is this V1-ing or V2-ing?", e.g., *is this grashing or speffing?* 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, although their responses were not recorded since the vocabulary questions only served as a distractor task.

Sentence comprehension distractor task: In this task, participants were presented with a sentence and had to identify what it meant by choosing one of two scenes displayed on the screen. Each of the two constructions occurred equally often within the set of comprehension questions. The verb was randomly selected among those attested with the construction in the input, but it was always different from the one presented in the following production question. The two scenes displayed the same action and the same two characters, but they differed in terms of the assignment of thematic roles (the agent in the first scene was the undergoer in the second scene, and *vice versa*). None of the two scenes had previously been presented in the exposure phase. 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 conducted after the production task was completed. It consisted of a standard acceptability judgment task. Participants were presented with 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 is showed in Figure 2.

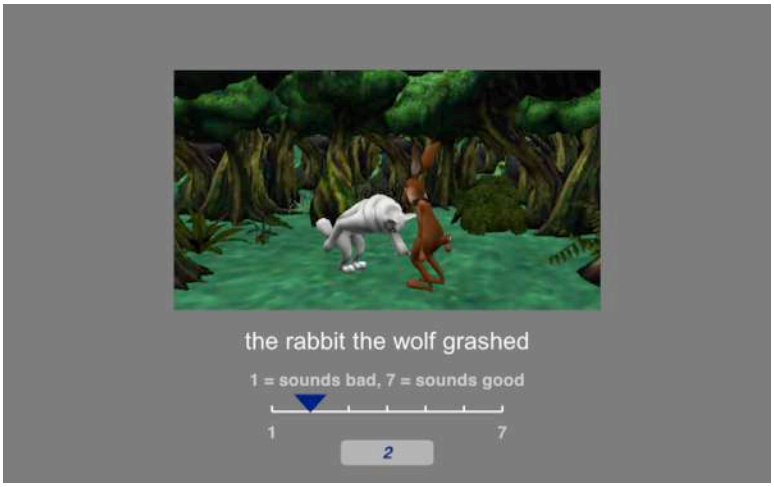


Figure 2: Example screenshot of the sentence rating task, with the verb *grash* ungrammatically used with OSV word order and an NP undergoer.

Participants provided responses on a 7-point Likert scale, with 1 being "sounds bad" and 7 being "sounds good." All six verbs shown during the exposure phase were used four times each (total of 24 items), once in each of the following kinds of sentences: grammatical SOV, grammatical ProSV, ungrammatical SOV (i.e., with a pronoun undergoer, e.g., *the panda him mooped*), ungrammatical ProSV (i.e., with an NP undergoer, e.g, *the pig the panda mooped* with *the panda* as agent). Since the last kind actually makes a possible string of words in the artificial language (i.e, it can be analyzed as an instance of SOV if no semantic information is provided), participants were explicitly instructed to pay attention to 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. We discuss the data collected on the first day subsequently (sections 2.4.2; 2.4.4). Immediately below, we describe the results of the production and sentence rating tasks in turn.[3] Our entire dataset is available as an online supplement.

2.4.1 Production task

The results of the production task were coded according to which construction was used. Sentences consisting of two noun phrases referring to the agent and undergoer arguments in that order, followed by the verb, were coded as "SOV"; sentences consisting of *him* followed by a noun phrase corresponding to the agent argument and by the verb were coded as "ProSV." Productions that did not fit either of these patterns were treated as errors and left out of the analysis. These accounted for two data points (0.5%) in the lexicalist condition and 33 (11%) in the alternating condition.[4] One production on day 2 in the alternating condition failed to be recorded due to technical issues. For SOV sentences, misnaming of one animal was ignored as long as the other animal was correctly named (thus allowing the thematic roles to be identifiable despite the error). In the event that the subject hesitated or produced multiple sentences, only their last full production was considered. As previously mentioned, the fact that the verb was provided insured that the correct novel verb was used. A few subjects (5/24 in the lexicalist condition and 3/18 in the alternating condition) failed to show evidence of having learned both constructions on day 2, insofar as they used the SOV construction in all of their productions and never used the ProSV construction. For the sake of completeness, we include these participants' data in the figures and model reported below, but the significant results are the same with or without their inclusion.

Before turning to the actual results we first present the predictions of both entirely lexically conservative behavior (Figure 3a) and entirely productive behavior (Figure 3b). In each figure, the hypothetical data are presented separately for SOV-only verbs, i.e., verbs that were attested only in the SOV construction in the input; and ProSV-only verbs, i.e., verbs that were attested only in the ProSV construction in the input. For each type of verb, productions in the two contexts are plotted separately: "NP-biased context", i.e., productions following the general question "what happened here?", and "Pro-biased context", i.e., productions following a question of the form "what happened to the <undergoer>?".

If participants strictly respect the verb-specific distribution evident during exposure, the context biasing questions used during the production task should not have

[3]As intended, by day 2, participants were at ceiling on the comprehension task. That is, they were successfully able to assign thematic roles to the arguments of the verbs, identifying the correct scene 97.6% of the time. As noted, we did not record accuracy on the vocabulary task for technical reasons, and because it was only used as a distractor task. Performance on this task could be expected to be close to ceiling since vocabulary was learned during the exposure phrase. In any case, the correct verbs were supplied to participants during the key production task.

[4]All 35 mistakes were of the same type: they correspond to uses of the OSV word order with an NP undergoer instead of a pronoun. Twenty of these mistakes were contributed by two participants in the alternating condition who failed to produce any pronouns at all. We return to these errors in the general discussion.

any effect on productions (Figure 3a). If, on the other hand, participants ignore the verb restrictions witnessed during exposure, and instead choose the construction that is best suited to the context, the results should pattern as in the hypothetical data in Figure 3b. In this case, productions depend entirely on which construction is better suited to the discourse.

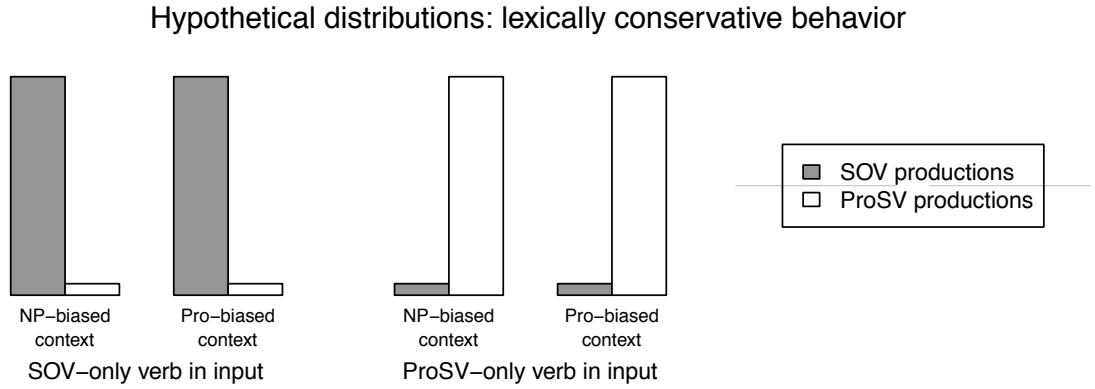


Figure 3a: Hypothetical proportions of SOV and ProSV productions predicted by lexical conservatism in which productions depend only on whether verbs were SOV or ProSV in the input.

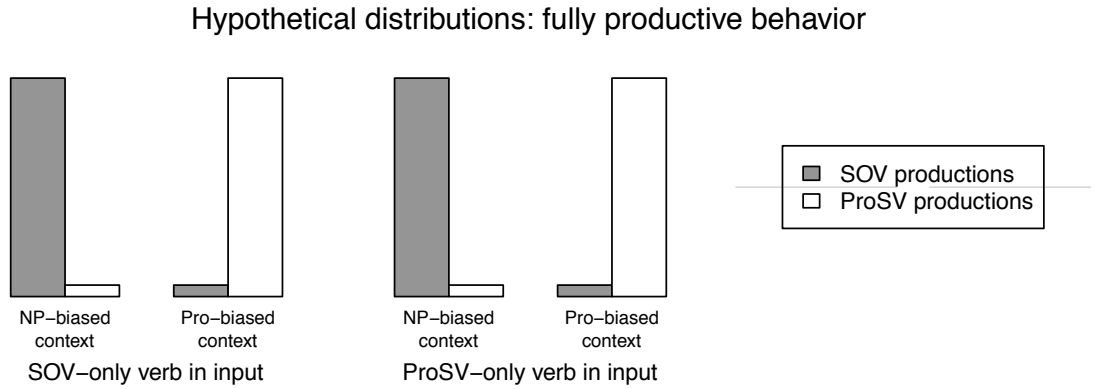


Figure 3b: Hypothetical proportions of SOV and ProSV productions predicted by full productivity, in which productions depend only on the information structure of the two constructions witnessed, and not on how the verbs were presented in the input.

We anticipated that the two verbs that had been witnessed alternating in the input (in the alternating condition only), and new novel verbs (i.e., verbs that did not occur during exposure) would both be used in whichever construction was more appropriate in the discourse context, since participants had no evidence of a lexical restriction favoring one construction over the other. As described below, this is what we found. Performance on new novel verbs serves as a type of baseline, since participants had no reason to be biased toward one construction or another.

The actual data are provided in Figure 4.

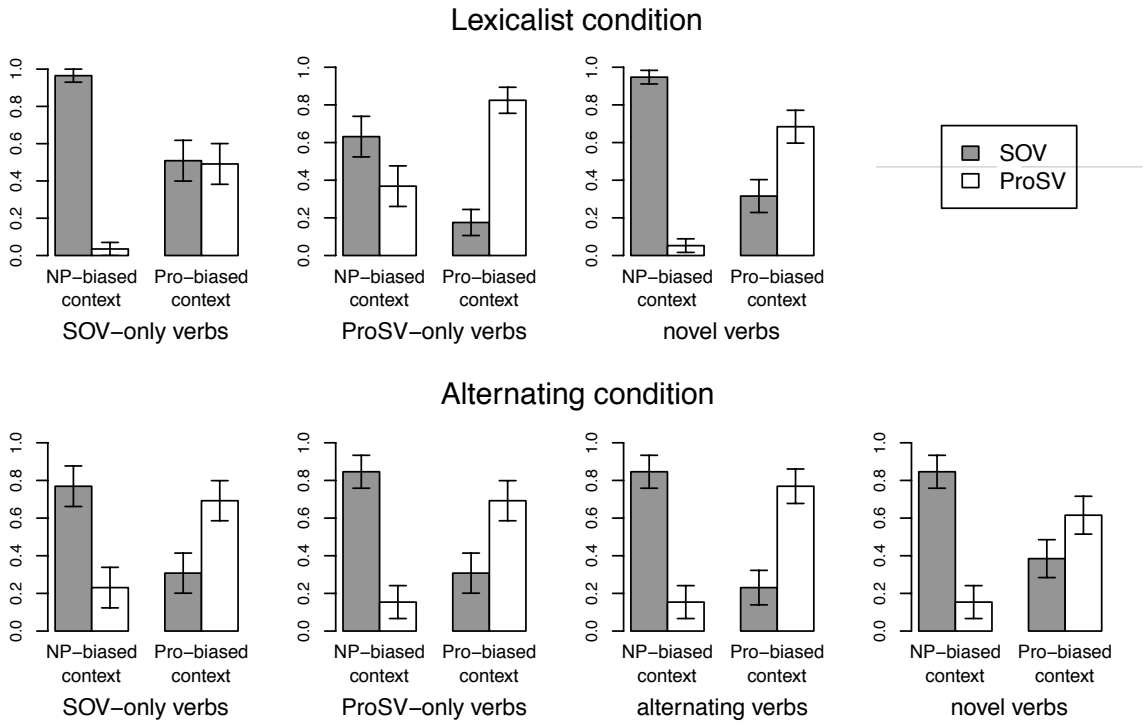


Figure 4: Proportions of SOV and ProSV productions in the lexicalist and alternating conditions, for each verb type and context type. NP-biased contexts refer to productions solicited by "what happened here?". Pro(noun)-biased contexts refer to productions solicited by "what happened to the <undergoer>?" SOV-only verbs were only witnessed with two lexical noun phrases and with the agent argument positioned before the undergoer. ProSV-only verbs were only witnessed with a pronominal undergoer positioned before a lexical agent. Alternating verbs (in the alternating condition only) were witnessed in both SOV and ProSV constructions. Novel verbs were previously unwitnessed (new) novel verbs.

We first detail the results descriptively before presenting the mixed effects logistic regression. There is an overall bias towards the SOV construction in both input conditions that we return to below. At the same time, it is clear that participants are sensitive to the functional difference between the two constructions. Specifically, in answer to the question, "What happened here?" (NP-biased contexts), there were more SOV productions when compared with answers to the question "What happened to the <animal>?" (i.e. Pro-biased contexts); conversely, there were more ProSV productions in the Pro-biased context than in an NP-biased context. The extent of participants' sensitivity to the functional difference, as demonstrated by their use of the construction that was more appropriate, varied according to the input condition.

In the alternating condition, even though only two of the six verbs were witnessed alternating, participants were very likely to use the construction that was the most appropriate in the discourse context, regardless of whether the verb had been previously witnessed in that construction or not. Verbs were produced in the SOV construction given an NP-biased context 83% of the time; they were produced in the ProSV construction given a Pro-biased context 63% of the time. In other words, participants were largely productive with both constructions, as in Figure 3b.

In the lexicalist condition, in which no verb was witnessed alternating, participants still showed some tendency to use the more appropriate construction, although this tendency was reduced when compared with the alternating condition. Hence, the results show a combination of lexical conservativeness, as in Figure 3a, and productive behavior, as in Figure 3b. Verbs witnessed only in the ProSV construction were used in the other, SOV construction, in a context that was biased for the SOV construction, 71% of the time. Verbs witnessed only in the SOV construction were used in the other, ProSV construction, in a context that was biased for a pronoun, 39% of the time. As expected, participants showed a tendency to use the two novel verbs in the contextually appropriate construction (i.e., 95% of uses of SOV in a NP-biased context, and 55% of uses of ProSV in a Pro-biased context).

To test for statistical significance, we submitted the data to mixed effects logistic regression, using the package lme4 in the R environment (Bates et al. 2011).[5] Each production of SOV or ProSV is one observation in the dataset. The dependent variable, SOV (binary), records whether the utterance used the SOV construction vs. the ProSV construction. We first focus on the non-alternating verbs (i.e., the verbs witnessed in SOV or ProSV only in the input), by testing for significant variation across conditions in the tendency for these verbs to be used in each construction (Table 1). We investigate the alternating verbs in the alternating condition in a separate analysis below (Table 2).

There are three predictors in the first model:

a) context bias (Bias), a binary variable that captures which biasing question was involved in the production (NP-biased question vs. Pro-biased question);
b) verb type (VerbType), a binary variable that captures whether a verb had been witnessed only in the SOV construction or only in the ProSV construction.
c) condition (Condition), a binary variable that records which input condition the participant was exposed to prior to the production task (alternating condition vs. lexicalist condition). In particular, this variable reflects whether the input included any alternating verbs (alternating condition) or not (lexicalist condition).

We performed standard model selection by first running the most complex model containing all interactions between fixed effects, and proceeding stepwise by removing non-significant interactions one by one (Baayen 2008). The final model contains Bias, VerbType, and Condition as main effects, and the interaction between VerbType and Condition. The fixed effects estimated by the fitted model are reported in Table 1.[6]

Random effects for subjects (Subject), verb forms (VerbForm), verb stems (VerbStem; i.e., verb forms without the morphological marker *-ee* or *-o*, when applicable), and verb meanings (Meaning) were included in the model in order to factor in subject-specific preferences and to control for potential constructional biases that might happen to be inherently associated with particular verb forms or meanings.

Following Barr et al.'s (2013) recommendations, we started with a maximal random effect structure containing random intercepts for participants and by-participant

[5]We used the 1.0-5 version of lme4. The *p*-values were calculated by the "summary" function from the package lmerTest version 2.0-3, which uses Satterthwaite's approximations (SAS Institute Inc., 1978).
[6] The full output of the lmer function for all mixed models discussed in this paper can be found in the online supplement.

random slopes for all other factors. The model initially failed to converge, and only did so when we removed all random slopes, thus only keeping random intercepts for the four factors. It should be noted that the variance of VerbForm, VerbStem, and Meaning is extremely small (< 0.0001), which means that these factors had virtually no effect on the subjects' productions. The same random effect structure was used for all models reported in this paper, on the basis of the same criteria. We found a classification accuracy (i.e., the percentage of data points for which the model predicts the right construction) of 79.04%,[7] which indicates that the model is a reasonably good fit for the data.

	Estimate	Std. error	z-value	p-value
(Intercept)	3.1838	0.3999	7.961	< 0.0001
Bias (Pro)	-2.3499	0.2732	-8.603	< 0.0001
VerbType (ProSV)	-1.3637	0.5118	-2.665	0.0077
Condition (alternating)	-1.8364	0.3286	-5.588	< 0.0001
VerbType (ProSV) : Condition (alternating)	2.0295	0.5424	3.741	0.0002

Table 1: Fixed effects of the logistic regression model predicting the production of the SOV construction, fitted to the data for SOV-only and ProSV-only verbs. Classification accuracy = 79.04%.

Since uses of the SOV construction were coded as '1', positive estimates of the fixed effects in Table 1 indicate that the corresponding factor has a positive impact on the use of the SOV construction. Conversely, negative values indicate that the factor favors the use of the ProSV construction. Results confirm the overall bias towards producing SOV, as evidenced by the significant positive intercept. This in line with previous observations that agent before undergoer is more natural for English speakers, possibly due to transfer effects from English, or due to a cross-linguistic bias of some sort (Boyd, Gottschalk, & Goldberg 2009; Wonnacott et al. 2008). At the same time, speakers tended to produce the construction that was most contextually appropriate; i.e., SOV productions were less common in contexts that were biased for a pronominal theme. This is reflected in the negative estimate for Bias (Pro). This context sensitivity is observed in both input conditions; it is not involved in a significant interaction with Condition.

SOV productions were less likely with verbs that had been witnessed only in the ProSV construction, an indication of lexical conservatism (i.e., the tendency to use a verb in the same construction with which it had been attested). This is reflected in the negative effect of VerbType (ProSV). Of particular interest is the fact that there is a significant interaction between VerbType and Condition, which means that the effect of verb type varies according to the input condition. Specifically, while verbs that had only been witnessed in the ProSV construction in the lexicalist condition were significantly less likely to be used in the SOV construction, the effect is largely eliminated in the alternating condition, when two of the six verbs were witnessed in both constructions. In

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

other words, the effect of lexical conservatism is specific to the lexicalist condition. Finally, the regression analysis also reveals a negative main effect of Condition (alternating), which means that subjects were slightly more likely to produce SOV in the lexicalist condition than in the alternating condition.

The results presented so far show that there are differences between the two groups of subjects in how they treated the non-alternating verbs. We now look more closely within the alternating condition in order to compare performance on the verbs that had been witnessed in both constructions with verbs that only appeared in one or the other construction during exposure, and with novel verbs. As noted above, subjects in the alternating condition appear to treat all verbs similarly, i.e., by combining them with the contextually appropriate construction. To quantify this, we submitted the data from the alternating condition to mixed effects logistic regression, with VerbType as well as Bias (and their interaction) as fixed effects, with SOV as the dependent variable, and with the same random effects as before. The fixed effects of this model are reported in Table 2.

	Estimate	Std. error	z-value	p-value
(Intercept)	1.3600	0.4499	3.023	0.0025
Bias (Pro)	-1.6767	0.5529	-3.033	0.0024
VerbType (alternating)	0.3992	0.6596	0.605	0.5451
VerbType (novel)	0.1591	0.6347	0.251	0.8021
VerbType (ProSV)	0.6429	0.6999	0.918	0.3583
Bias (Pro) : VerbType (alternating)	-1.2212	0.8508	-1.435	0.1512
Bias (Pro) : VerbType (novel)	-0.1557	0.8083	-0.193	0.8472
Bias (Pro) : VerbType (ProSV)	-0.8371	0.8675	-0.965	0.3346

Table 2: Fixed effects of the logistic regression model predicting the production of the SOV construction in the alternating condition. Classification accuracy = 75.2%.

Results demonstrate that subjects in the alternating condition did generalize all verbs to both constructions equally, according to the context of utterance. This tendency did not significantly vary according to whether the verbs were witnessed exclusively in either construction (SOV-only and ProSV-only verbs), in both constructions (alternating verbs) or in none of the constructions (novel verbs). That is, as shown in Table 2, there is a negative main effect of Bias (Pro), and this effect is not involved in any significant interaction with the different levels of VerbType. This strengthens our general point that participants were fully productive in the alternating condition.[8]

2.4.2 Performance on production task day 1

Performance after one day of exposure contained a fair number of errors indicating that, as anticipated, participants had not yet fully learned the two constructions. Productions that were neither SOV nor ProSV included 33 data points (9%) in the lexicalist condition

[8] This same analysis cannot be done on the lexicalist condition data because there were no alternating verbs in that condition. In the lexicalist condition, since there was an effect of how individual verbs had been witnessed in the input (verb type), it is also clear that participants were not maximally productive in their choice of construction.

and 57 (20%) in the alternating condition. Five productions on day 1 in the lexicalist condition failed to be recorded due to technical issues.

We ran the same mixed effects model reported in Table 1 on the data from the first day, and found that participants already displayed a significant effect of context (which question was asked). They also displayed evidence of lexical bias but without the interaction found on day 2; i.e., the lexical bias remained in evidence in the alternating condition. This suggests that participants were more willing to move beyond the lexically specific input in order to use the constructions in discourse appropriate ways after the second day of exposure.

We also ran the same mixed effects model reported in Table 2 on the data from day 1. As on day 2, we found that participants in the alternating condition already had a general tendency to use the contextually appropriate construction (as shown by a main effect of Bias), and there were no significant interactions with the different levels of VerbType. At the same time, participants were more likely to use SOV with SOV-only verbs than with alternating verbs, although this lexical conservatism was not in evidence with the ProSV-only verbs. As previously mentioned, this tendency for a degree of lexical conservatism is no longer found on day 2 in this alternating group. Overall, participants displayed more sensitivity to context on day 2.

For reasons of space, we do not report the full models here, but they can be found in the online supplement.

2.4.3 Sentence rating task

The data from 40 subjects in the sentence rating task were considered in the analysis. Two subjects in the alternating condition were excluded because they did not recognize any functional difference between the constructions: i.e., in the production task, they systematically produced the OSV word order with a lexical NP undergoer (and never with a pronoun) in combination with ProSV-only verbs (but never with SOV-only verbs), and in the post-experiment debriefing, they did not mention that the OSV word order was restricted to pronoun undergoers, contrary to all other participants. Their failure to properly acquire the ProSV construction prevents their data from being comparable with those from the rest of the group. Interestingly, the fact that these subjects disregarded the information structure difference between constructions makes their learning task comparable to that given by Wonnacott et al. (2008) in their first experiment (cf. Section 1), and, accordingly, led them to be lexically conservative.

As is standard practice when handling grammaticality rating data, we normalized the raw ratings given on the 7-point scale to z-scores in order to control for the fact that subjects often use the scale in different ways. Z-scores are calculated according to the following formula: $z = (r - \mu) / \sigma$, where r is the raw 7-point-scale rating, μ is the mean of all ratings provided by the same participant, and σ is the standard deviation of these ratings. That is, the conversion to z-scores consists in replacing each rating by a value that indicates by how many standard deviations it diverges from the subject's average rating.

Figure 5 presents the distributions of z-scores in the two input conditions and for each verb type, in the form of box plots. The distribution of each construction is plotted separately: the two attested constructions first, SOV (e.g., *the panda*$_{agent}$ *the cat*$_{undergoer}$ *glimmed*) and ProSV (e.g., *him*$_{undergoer}$ *the panda*$_{agent}$ *glimmed*), followed by two

unattested constructions, SProV (e.g., **the panda$_{agent}$ him$_{undergoer}$ glimmed*) and OSV (e.g., **the cat$_{undergoer}$ the panda$_{agent}$ glimmed*). In these diagrams, 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). The values outside this range are outliers and represented by bullets in the plots.

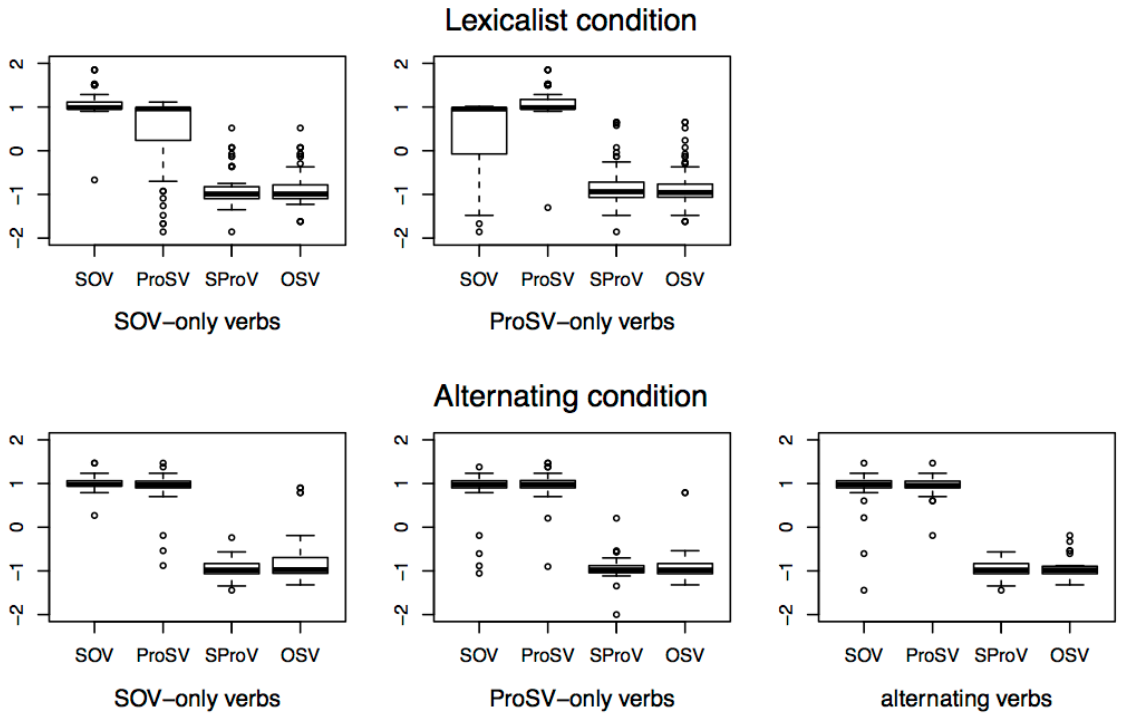


Figure 5: Box plots of the distribution of grammaticality ratings (z-scores) provided by participants in the lexicalist condition (top) and in the condition in which one third of verbs alternated (bottom), for each verb type and each construction.

As can be seen in the box plots, the sentence ratings are largely polarized, in that they mostly occupy the extreme ends of the standardized scale. Overall, instances of the two constructions that had been witnessed during the exposure phrase (the two boxes on the left of each plot) were judged to be markedly more acceptable than instances of unattested constructions (the two boxes on the right of each plot). That is, for all verb types in all conditions, there is a clear divide between attested constructions, which are generally considered acceptable, and the unattested constructions, which are judged roughly equally unacceptable.

Focusing more specifically on differences between verb types in the attested constructions, we observe that, as in the production task, subjects in the alternating condition, in which one third of the verbs alternated, do not differentiate among verbs, as they judge them all fully grammatical in both SOV and ProSV. However, in the lexicalist

condition, two rating distributions, while still centered on positive scores, contain a long tail of values ranging towards the middle of the scale. These distributions correspond to ProSV sentences with SOV-only verbs, and SOV sentences with ProSV-only verbs, which many subjects rated as less acceptable than the same sentences with ProSV-only verbs and SOV-only verbs respectively. This again reflects some tendency toward lexical conservatism in the lexicalist condition: for a reasonable number of subjects, sentences combining a verb with a construction it has not been attested in are considered less grammatical than sentences combining a verb with its usual construction. Note, however, that such sentences are still not considered as bad as sentences with unattested constructions, which, regardless of the verb, strike all subjects as clearly ill-formed.

To test whether these differences are significant, we submitted the ratings of sentences with SOV-only verbs and ProSV-only verbs used in the attested constructions (SOV and ProSV) to mixed effects linear regression. The regression model contains two predictors: (i) Match, a binary variable which records whether the verb is used in the same construction that it has been exclusively witnessed in, and (ii) Condition, a binary variable that records which input condition the participant was exposed to (as in Section 2.3.1). As previously, by-subject, by-verb-form, by-verb-stem, and by-meaning random effects were also included, but only by-subject effects captured significant variance.

The fixed effects of the linear model are reported in Table 3. Both predictors and their interaction are significant. There is a positive main effect of Condition, showing that sentences are overall judged more grammatical by participants in the alternating condition than by participants in the lexicalist condition. In the lexicalist condition, sentences are judged more acceptable when they contained a verb used in the construction with which it had been witnessed in the input. This is evident in the positive main effect of Match, and the negative interaction with Condition (alternating), which makes the effect of Match mostly specific to the lexicalist condition.

	Estimate	Std. error	t-value	p-value
(Intercept)	0.55083	0.05741	9.595	< 0.0001
Condition (alternating)	0.26614	0.09795	2.717	0.0079
Match (true)	0.53198	0.06013	8.847	< 0.0001
Condition (alternating) : Match (true)	-0.38381	0.10840	-3.541	0.0004

Table 3: Fixed effects of the linear regression model predicting the z-score ratings provided by subjects in the sentence rating task.

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

2.4.4 Performance on judgment task day 1

We also ran the same mixed model on the sentence rating data from day 1; for reasons of space, we do not report the full model here, but it can be found in the online supplement. We found that only the positive main effect of Match (true) was significant (although smaller); importantly, the interaction of this predictor with Condition (alternating) was not significant. This means that, on day 1, both groups tended to judge sentences to be

more acceptable when they contained a verb that was witnessed with the same construction in the input. In other words, the lexical bias was in evidence in both conditions on day 1, similarly to what we reported earlier for the production task.

2.5 Discussion

The present experiment exposed participants to two constructions that differed in terms of information structure as well as word order. The SOV construction involved lexical noun phrases and an <agent-undergoer-verb> order, whereas the ProSV construction expressed the undergoer argument with a pronoun in an <undergoer-agent-verb> order. During the production task, the questions used to elicit responses provided contexts in which the undergoer argument was either new or already given in the discourse, making the function of one or the other construction better suited to the discourse demands.

In the alternating condition, when only two out of six verbs were witnessed alternating, participants readily extended all verbs for use in the contextually appropriate construction. That is, participants tended to produce all of the verbs in the more appropriate construction regardless of whether the verb had been attested in that construction or not. Participants also tended to judge all verbs as acceptable in either construction, regardless of whether the verb had been witnessed in both constructions or not. These findings stand in contrast to the results of Wonnacott et al. (2008)'s second experiment; recall that they had found that, when only a third of the verbs were witnessed alternating, participants were quite lexically conservative.

At the same time, and in line with Wonnacott et al. (2008), a comparison of our lexicalist and alternating conditions demonstrates that the statistics in the exposure also matter: participants in the lexicalist condition, in which no verbs were witnessed alternating, were more likely to be lexically conservative in both their productions and in their judgments. Hence, Wonnacott et al.'s observation that "learners are sensitive to statistical information above the level of individual verbs" (p. 204) carries over to our more ecologically valid case of a language containing two constructions with a variation in pragmatic function. At the same time, in the present experiment, even in the lexicalist condition, lexical conservativeness competes with an appreciable tendency of subjects to apply the context-driven alternation to all verbs. In other words, the behavior of participants in the lexicalist condition lies somewhere between the two hypothetical distributions shown in Figures 3a and 3b, as it presents aspects of both lexical conservativeness and constructional productivity.

A key difference between Wonnacott et al.'s (2008) study and the present one is that the two constructions used in the present study differed systematically in terms of information structure. Results demonstrated that participants were sensitive to the information structure differences in that both groups displayed a tendency to use each construction in the appropriate discourse contexts. Moreover, participants did not simply use pronouns freely. They only used pronouns in the ProSV construction; there were no SProV productions. Thus it is clear that participants assigned information structure restrictions to the distinct word order constructions.

The recognition of different information structure properties may have led participants to infer that the distinction between the two constructions was not additionally conditioned by verb semantics or phonological form. In the absence of any functional distinction between two constructions, the participants in Wonnacott et al.'s

study were much more willing to attribute the difference between the two constructions to verb-level stipulations.

While our experiment was similar in idea and purpose to Wonnacott et al.'s, we used a different procedure, a different subject population, different word order constructions, fewer input sentences, and two experimental sessions instead of five. It is possible that our participants were more likely to generalize on the basis of the information structure of the constructions instead of being lexically conservative, simply because they could not keep track of the preferences of individual verbs or did not detect verbs' lexically-specific behavior. Results after the initial day of exposure makes this somewhat unlikely since participants were more lexically conservative than on the second day. Nonetheless, in order to investigate this possibility directly, we tested a new set of participants in a second experiment in which the two constructions were functionally equivalent.

3. Experiment 2

In Experiment 2, we test participants with a third version of the artificial language that was similar to Wonnacott et al.'s (2008) lexicalist language; there were two functionally indistinguishable constructions that were instantiated by two distinct sets of verbs. We hypothesize that participants would attempt to keep track of the syntactic behavior of each verb, and as a result be predominantly lexically conservative in this case, replicating the results of Wonnacott et al.'s first experiment. If we find that participants use verbs haphazardly in either word order construction, this could mean that they were unable to memorize the syntactic behavior of each verb with the type and amount of input provided. If, however, we find that they consistently use verbs in the only construction with which they were witnessed, this would indicate that participants in Experiment 1 were fully *capable* of learning the lexically-specific behavior of each verb. This finding would demonstrate that the tendency to generalize in both conditions of Experiment 1 is not attributable to a simple failure to detect or remember the way each of the six novel verbs was used during the exposure phase.

3.1 Participants

The participants were 12 undergraduate students at the department of psychology at Princeton University (6 female, 6 male, aged 18-22, mean 19.6) who participated in the study for course credit. All were native speakers of English and had normal or corrected vision.

3.2 Materials

The artificial language used in Experiment 2 is similar in all respects to that used in Experiment 1, except that it does not contain any pronouns. Hence, the functional distinction between the two word order constructions is lost, as both are consistently used with two full lexical noun phrases. These two constructions will simply be referred to as SOV (e.g., *the panda*$_{agent}$ *the monkey*$_{undergoer}$ *pilked*) and OSV (e.g., *the wolf*$_{undergoer}$ *the rabbit*$_{agent}$ *mooped*).

The exposure sets given to participants contained the same number of sentences and verbs as in Experiment 1. The assignment of verbs to distributional classes was identical to the lexicalist condition of Experiment 1: three verbs occurred exclusively in SOV, and the other three verbs occurred exclusively in OSV. The questions used at test

were also identical to those used in Experiment 1, as were all other details regarding the artificial language and the exposure set.

3.3 Procedure

The procedure was identical to the one used in Experiment 1. Although the pronoun-biased context should not elicit a different response here since neither of the constructions is used with pronouns in the input, we retained the variation in the questions asked in the production task (i.e., "what happened here?" vs. "what happened to the <undergoer>?"). This was done in order to make the results maximally comparable with those of Experiment 1, but the question manipulation did not play any role in Experiment 2. For the same reason, the stimuli in the sentence rating task also contained pronouns, even though the input given to subjects did not.

3.4 Results

3.4.1 Production task

We used the same coding scheme as in Experiment 1.[9] The proportions of SOV and OSV constructions in the subjects' productions with each kind of verb are diagrammed in Figure 6.

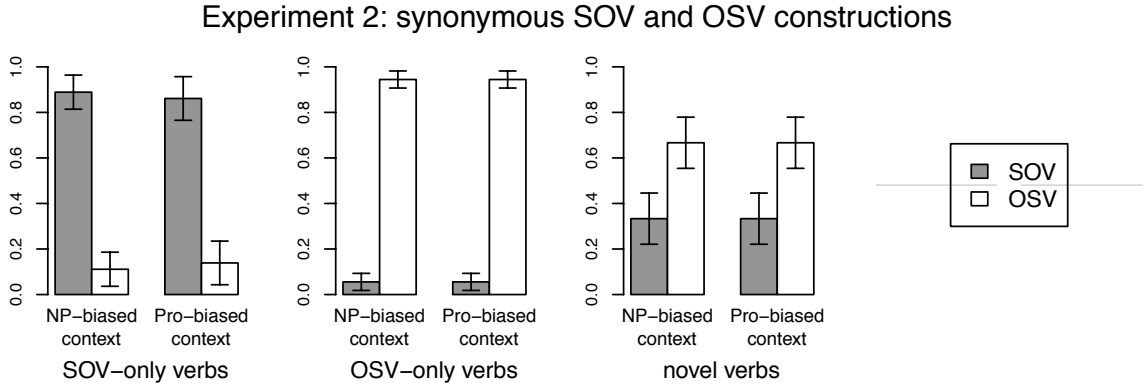


Figure 6: Proportions of SOV and OSV productions in Experiment 2 (involving equivalent constructions) for verbs presented only in the SOV construction and for verbs presented only in the OSV construction. No pronouns were witnessed during exposure and none were produced at test.

Results demonstrate a strong tendency towards lexically conservative behavior, replicating the finding of Wonnacott et al. (2008)'s Experiment 1, and demonstrating that participants are fully capable of learning verb-specific distributional information with the amount of exposure provided in both present Experiments 1 and 2. That is, participants overwhelmingly produced the SOV construction with verbs that had been witnessed in that construction and the OSV construction with verbs that had been witnessed in that construction. Their productions directly mirror the input. As expected, the biasing questions did not have any effect: participants produced each construction in the same proportion in either a NP-biased context or a Pro-biased context.

[9]Performance in the comprehension task was comparable to that of Experiment 1. Participants identified the correct scene 94% of the time on average.

For the new novel verbs that had not been witnessed in either construction, there seems to be a numerical preference for the OSV word order (about 66% of all productions). This numerical trend towards OSV runs counter to the bias toward SOV in Experiment 1 and results from previous studies (e.g., Boyd et al. 2009; Wonnacott et al. 2008). The difference is, however, not significant (paired $t(11) = 1.62$, $p = 0.13$). To check whether the lack of significance was simply due to the small sample size (12 participants, 4 observations per subject), we ran a power analysis following Cohen (1988).[10] We found that the power of our *t*-test (i.e., the probability of finding an effect that is there: 1 - Type II error) is reasonably high (*Power* = 0.83; $n = 12$; Cohen's $d = 0.93$; significance level = 0.05), and at any rate above the 0.8 threshold recommended by Cohen (1988). Hence, it seems that our sample of twelve subjects should be large enough to detect a significant effect of this size. We can thus conclude to the absence of a real bias towards OSV order for novel verbs in Experiment 2.

At any rate, there might be a circumstantial explanation as to why we did not find any bias towards SOV. When asked to explain how they used the new novel verbs, many participants reported that they looked for semantic or formal similarities with verbs attested in the exposure set, and applied the corresponding construction. Hence, the slightly higher frequency of OSV in the productions with novel verbs might merely reflect fortuitous ways in which the novel verbs were semantically or phonologically related to those that were randomly selected for inclusion in the exposure set, rather than an inherent preference for a particular word order. Because the bias toward SOV word order in Experiment 1 was not in evidence in Experiment 2, we leave this aspect of our results aside for future investigation.

Recall that the difference between the exposure in the lexicalist condition of Experiment 1 and the exposure in Experiment 2 is that the former used pronouns to distinguish the functions of the two witnessed constructions: ProSV treated the undergoer argument as topical, while SOV treated the undergoer as new to the discourse. In Experiment 2, no pronouns were used; both constructions involved two lexical NPs and thus were functionally equivalent. We therefore did not expect participants to use one construction over the other after either type of context-biasing question, and results demonstrate that they did not. Instead, participants in Experiment 2 were strongly lexically conservative, using each verb in the construction in which it had been witnessed during exposure.

To quantify whether the lexical conservatism evident in Experiment 2 was appreciably stronger than that in the lexicalist condition of Experiment 1, we compared the results of these two groups (without the novel verbs). We submitted this dataset to mixed effects logistic regression, as in Section 2.3.1, except that the predictor Condition was replaced by the binary variable Equivalent which corresponds to the distinction between Experiment 2 (Equivalent (true)) and the lexicalist condition of Experiment 1 (Equivalent (false)). Also note that the verbs are divided into SOV-only verbs and OSV-only verbs; the latter corresponds to ProSV-only verbs in the Equivalent (false) group. As before, the model was selected by initially including all possible interactions among the three factors of interest (Bias, Equivalent, and VerbType) and incrementally removing non-significant interactions. The fixed effects of the final model are summarized in Table 4.

[10]We used the "effsize" R package to calculate Cohen's *d* and the "pwr" package to calculate power.

	Estimate	Std. error	z-value	p-value
(Intercept)	3.9066	0.6273	6.227	< 0.0001
Bias (Pro)	-2.8129	0.3954	-7.115	< 0.0001
VerbType (OSV)	-2.2545	0.3938	-5.725	< 0.0001
Equivalent (true)	-0.5146	1.1023	-0.467	0.6406
Equivalent (true) : Bias (Pro)	2.5683	0.8611	2.982	0.0029
Equivalent (true) : VerbType (OSV)	-5.1771	1.1469	-4.514	< 0.0001

Table 4: Fixed effects of the logistic regression model predicting the occurrence of the SOV construction, fitted to the data for SOV and OSV verbs from Experiment 2 and the lexicalist condition of Experiment 1. Classification accuracy = 87.21%.

The main effect of group, Equivalent (true), is not significant, but the variable is involved in two highly significant interactions, showing that the effects of the other factors vary substantially between groups. First, Bias has virtually no effect in the Equivalent group, whereas subjects tended to produce fewer SOV sentences in a Pro-biasing context in the non-equivalent group. That is, participants were entirely lexically conservative only when the constructions were functionally equivalent (the negative effect of Bias (Pro) is countered by its positive interaction with Equivalent (true)). Thus the effect of the discourse-biasing context questions evident in the lexicalist condition in Experiment 1 is not found in Experiment 2. Second, there is an effect of lexical conservatism in both groups, but this effect is markedly more pronounced when the constructions were equivalent (Experiment 2). That is, subjects tended to produce more OSV sentences with OSV-only verbs overall, and were even more likely to do so when the two constructions were equivalent: i.e., VerbType (OSV) and its strong interaction with Equivalent (true) are negative.

3.4.2 Sentence rating task

The results of the acceptability rating task for verbs witnessed in the SOV construction and those witnessed in the OSV construction are presented in Figure 7 in the form of box plots for each construction. As before, we first standardized the raw 7-point-scale ratings for each subject by converting them into z-scores. As can be seen in the diagram, sentences with verbs of each class are considered acceptable by participants only when they instantiate the word order construction with which the verb has been attested in the input. Interestingly, replacement of the undergoer argument by a pronoun (viz. SProV instead of SOV, or ProSV instead of OSV) triggered lower acceptability ratings, which was likely due to the lack of familiarity with pronoun use in the input. At the same time, the use of pronouns in the correct order does not lower acceptability as much as shifts in word order do (viz. OSV instead of SOV or vice-versa); the latter are towards the bottom end of the acceptability scale.

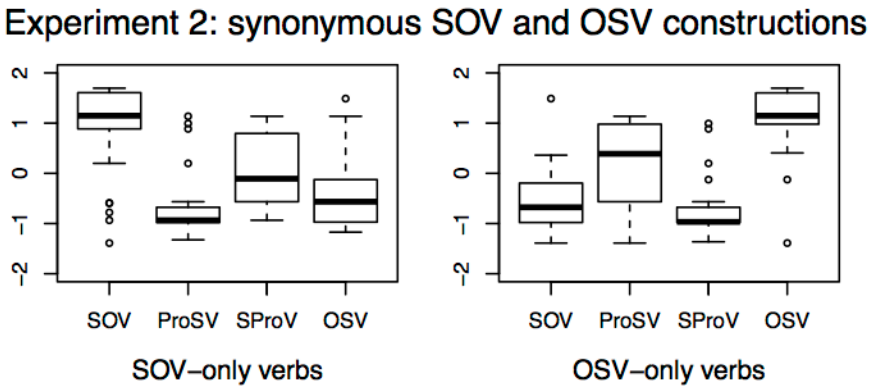


Figure 7: Experiment 2. Box plots of grammaticality ratings (z-scores) for four sentence types (SOV, ProSV, SProV and OSV) with verbs that had been presented only in SOV during exposure (left) or only in OSV during exposure (right). SOV and OSV were functionally equivalent.

To quantify the effect of constructional equivalence on grammaticality ratings, we combined the sentence rating task data from Experiment 2 (involving equivalent constructions) with those from the lexicalist condition of Experiment 1 (non-equivalent constructions), excluding the new novel verbs. We submitted this dataset to mixed effects linear regression analysis, along the same lines as in Experiment 1 (cf. Section 2.3.2). The regression model contains the binary variables Equivalent and Match as fixed effects, and by-subject random effects; as in Experiment 1, Match codes whether the verb in the rated sentence is used in the construction it has been witnessed with. As previously, we converted the 7-point-scale ratings into z-scores. The fixed effects of the model are reported in Table 5.

	Estimate	Std. error	t-value	p-value
(Intercept)	0.55083	0.0703	7.836	< 0.0001
Equivalent (true)	-0.99591	0.12176	-8.180	< 0.0001
Match (true)	0.53198	0.06969	7.634	< 0.0001
Equivalent (true) : Match (true)	0.98389	0.12070	8.152	< 0.0001

Table 5: Fixed effects of the linear regression model predicting the z-score ratings provided by subjects in the sentence rating task.

As seen in Table 5, both fixed effects and their interaction are highly significant. Ratings in the group with equivalent constructions were on the whole less positive than in the group with functionally distinct constructions: there is a negative main effect of Equivalent (true). Ratings were better overall when the verb was used in the same construction it had been witnessed with in the input: there is a positive main effect of Match. Of most relevance is the fact that, while an effect of lexical conservatism on grammaticality ratings exists in both conditions, it is much stronger when the two constructions are equivalent than when they exhibit a difference in information structure. This is evidenced by the strong positive interaction between Match and Equivalent.

4. General discussion

When, and on what basis, do learners generalize a construction beyond their exposure for use with words that have not been witnessed in that construction? This is a question that

has bedeviled researchers of argument structure for quite a long time (e.g., Lakoff 1970; Braine 1971; Baker 1979; Bowerman 1988; Pinker 1989; Goldberg 1995; Ambridge et al. 2014; Ambridge et al. 2008). Recent work on artificial grammar learning has demonstrated that the statistics in the input can play a key role. The more verbs that are witnessed alternating between two constructions, the more learners are willing to assume that other verbs can alternate as well (Perek 2015: Ch. 7; Wonnacott et al. 2008; Wonnacott 2011). The present findings confirm that fact, but temper the conclusion by demonstrating that learners do not produce language "blindly" on the basis of some predetermined aspect of the statistics in the input (cf. also Schumacher, Pierrehumbert, & LaShell 2014). Instead, learners critically and appropriately take the function of constructions into account in determining whether to generalize to new instances.

In Experiment 1, two constructions were distinguished on the basis of their word order and their functions: certain discourse contexts were more appropriate for one construction than the other. In particular, one word order construction included a pronominal undergoer ($Pronoun_{Undergoer}$ NP_{Agent} V: the "ProSV" construction), and a different word order construction contained full two lexical NPs (NP_{Agent} $NP_{Undergoer}$ V: the "SOV" construction). One or the other construction was made more appropriate by means of asking during the production task, either "What happened here?" or "What happened to the <undergoer>?"

In the alternating condition, even though only a minority of verbs were witnessed in both constructions, learners readily produced whichever construction was more appropriate in the discourse, regardless of how the verb had been witnessed during exposure. Acceptability judgments confirmed that learners ignored the lexical biases present during exposure, and readily generalized each construction for use with all verbs.

Moreover, even in the lexicalist condition, in which each verb was uniformly witnessed only in a single construction, participants still showed some tendency to generalize beyond their input and use verbs in the other construction when the other construction was more appropriate. While this group showed more sensitivity to the lexical biases in the input, the majority of participants still judged verbs to be reasonably acceptable when used in the alternative construction.

The second experiment ruled out the possibility that participants were simply unable to keep track of the lexical biases in the input. That is, when the two constructions witnessed in the input were functionally identical, participants uniformly displayed lexical conservatism in both their productions and in their acceptability judgments, as had been found previously (Wonnacott et al. 2008). The present findings raise the question of whether learners' conservative behavior in the second experiment and in Wonnacott et al. (2008) was a result of a type of indirect negative evidence, namely, *statistical preemption* (Goldberg 1995). Statistical preemption is the process whereby speakers learn to avoid using a verb in a particular construction, CxB, if the verb is consistently witnessed in a competing construction, CxA, instead (cf. Ambridge et al. 2014; Boyd & Goldberg 2011; Brooks & Tomasello 1999; Goldberg 2006; 2011; Payne et al. 2013; Perfors et al. 2010; Poser 1992; Robenalt & Goldberg, to appear). When two constructions have identical functions, hearing a verb in one construction is tantamount to not hearing it in the other construction. When two constructions are distinguishable on the basis of information structure, participants show a tendency to generalize on this basis, giving less weight to lexical differences in the exposure than they do when the two constructions serve

identical functions. The present results suggest that participants are willing to use verbs in new ways when such an extension is better suited to the demands of the discourse. Witnessing even a minority of verbs alternating led participants to readily extend all verbs for use in either construction as conditioned by the appropriate discourse context.

Our adult participants likely know implicitly that pronominal use is not typically conditioned by verbs. Yet in Experiment 2, even though half of the questions asked during the test phase made the undergoer argument topical and thus in principle available for being referred to with a pronoun, participants did not spontaneously produce any pronouns, since none were witnessed during exposure. This indicates that speakers were quite sensitive to the presence of pronouns in Experiment 1. Most importantly, Experiment 2 demonstrates that participants were capable of learning verb-specific restrictions with the amount and type of input they were given. When there was no conflict between verb-specific restrictions and the function of the two constructions, participants were entirely lexically conservative, in a replication of Wonnacott et al. (2008).

Our central focus is on participants' productions and judgments in Experiment 1, where verb-specific input was in competition with the discourse functions of grammatical constructions. Participants in Experiment 1 *could have* respected both verb biases *and* contextual biases by simply substituting a pronominal argument for an NP argument, or vice versa. That is, they could have used SProV or SOV for SOV-verbs; and ProSV or OSV for OSV-verbs. However, none of the participants showed any tendency toward doing this; the few participants who produced sentences that were neither instances of the SOV nor the ProSV construction produced OSV utterances regardless of discourse context (these were described as errors above).

Instead, participants learned two distinct word-order constructions and associated them with distinct discourse functions, presumably on the basis of their prior knowledge that pronouns are used to refer to discourse-given and topical arguments of verbs. Accordingly, participants clearly showed context sensitivity at test, using the construction that was most appropriate according to their existing discourse biases. Moreover, they used pronouns only in the ProSV (<undergoer-agent-verb>) construction: no participant ever produced SProV. And participants displayed a keen sensitivity to the different discourse functions of the constructions; even in the lexicalist condition, participants displayed a tendency to use ProSV when the undergoer argument was topical and to use SOV (<agent-undergoer-verb>) when it was not. In the alternating condition, in which (only) two out of six verbs were witnessed in both constructions, participants completely ignored the apparent verb-specific restrictions. That is, participants regularly *overrode* evidence of verb-specific word order constraints in favor of using whichever construction better suited the demands of the discourse. In essence, participants—even in the lexicalist condition but especially in the alternating condition—recognized that the constructions were "about" information structure and not about the choice of individual verbs, and they generalized accordingly.

The fact that learners generalized on the basis of information structure in Experiment 1, largely disregarding the role of verb-specific biases, suggests that the perceived functions of constructions play an important role in determining which dimensions are relevant to generalization. It is possible that the adult learners in our experiment were able to hone in on information structure as a relevant dimension in part

due to the fact they had already learned a number of English constructions that serve the function of packaging information in various ways, e.g., passives, topicalization, left dislocation, and clefts.

In natural language, the learners' task is highly complex, as constructions are typically conditioned by multiple factors. Many argument structure constructions are conditioned by lexical semantics, information structure, and even phonology. For example, information structure plays a role in the distinction between the double object and *to*-dative constructions, as already mentioned (e.g., Thompson 1990; Bresnan et al. 2007), as does verb semantics (Green 1974; Pinker 1989; Goldberg 1995; Ambridge et al. 2008), and phonology (Green 1974; Gropen et al. 1989; Ambridge et al. 2012).

In fact, the functions of individual constructions, if defined generally, can supply a multitude of relevant dimensions by which slots, whether verbal or nominal, may be generalized. For example, the rhetorical question made popular in America by a dairy commercial, "Got milk?" has been generalized to a variety of other nouns as in (1) (found on T-shirts at http://www.cafepress.com/aboriginalz/2987220 or as decals at http://picclick.com/got-jesus-Vinyl-Decal-7-x-25-11-130447346925.html):

(1) a. Got soccer?
b. Got Jesus?
c. Got hope?

Normally, "milk," "soccer," "Jesus," and "hope" are not considered members of the same category, but the phrasal construction as a whole picks out a relevant dimension that includes these instances. In particular, the construction is used to rhetorically ask whether the listener or reader has a certain essential component of what the speaker construes to be a good life. The examples in (2) are all quite pragmatically odd because it is difficult to imagine contexts in which sour milk, despair, or depression are construable as essential to well being.

(2) a. ?? Got sour milk?
b. ?? Got despair?
c. ?? Got depression?

The fact that different constructions can be conditioned by a wide range of factors, and combinations of factors, presents a deep challenge to research in language acquisition. How do speakers come to recognize which dimensions are relevant for generalizing a given construction? The present results suggest that generalization depends at least in part on learners identifying the function of an individual construction and then generalizing on the basis of factors that are relevant to that function.

Given the complexity of the task, it seems likely that children may well display less facility in identifying the functions of individual constructions, and therefore may be less well attuned to relevant dimensions for the sake of generalization. This may result in more conservative behavior when new constructions are being learned (see e.g., Tomasello 1992; 2000; Akhtar 1999; Boyd & Goldberg 2012). Alternatively, less facility in determining which dimensions are relevant to a given construction may result in children generalizing on the basis of dimensions that are not relevant to adults, leading to

overgeneralizations or even language change. Future work with children at various ages will be important to investigating these issues.

The present study provides evidence that learners do not completely ignore distributional information, even if it is not relevant to a construction's function. In the lexicalist condition of Experiment 1, learners showed some tendency to be lexically conservative even though the function of the constructions was unrelated to verb semantics. A real world example of a functionally irrelevant dimension playing some role in generalizability is the fact that the phonology of verbs is relevant to which verbs occur in the double-object construction (Green 1974; Gropen et al. 1989; Ambridge et al. 2012), even though phonology is presumably not relevant to the construction's function. We conjecture that factors that are *relevant* are more highly weighted, but speakers are capable of taking additional factors into account in determining which construction to use, with which verb, in a given context.

5. Conclusion

In line with previous findings (Wonnacott et al. 2008), the present results demonstrate that both language-wide and item-specific statistics play a role in language learning. In fact, much previous work on argument structure acquisition has emphasized its input-driven and verb-specific nature (e.g., Tomasello 1992; 2000; Akhtar 1999; Levin 1993). The present paper demonstrates the function of the constructions involved play a critical role as well. Learners are willing to use verbs in unwitnessed ways when the function of the target construction is better suited to the discourse context than a witnessed construction would have been. Speakers are in fact quite likely to use a verb in a second construction that had not been previously witnessed with that verb, if they have witnessed even a minority of verbs appearing in both constructions. Judgments of acceptability confirmed learners' willingness to use verbs in unwitnessed ways. Thus while the role of statistical distribution is critical, as has been amply demonstrated by a wide range of work and is confirmed by the comparison of lexicalist and alternating conditions in the present study, learners nonetheless display a striking willingness to go beyond the statistical distribution of their exposure in order to use a construction with a new verb, when the discourse context warrants it.

Acknowledgements

We are grateful to Stefan Gries, Clarice Robenalt, and Laura Suttle for statistical advice, as well as to Martin Pickering and two anonymous reviewers of an earlier draft for very helpful comments and suggestions. This research was supported by a postdoc scholarship granted to the first author by the DAAD (German Academic Exchange Service), and by an Einstein visiting fellowship from the Einstein Foundation in Berlin to the second author.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, *54*(4), 541-553.

Fisher, C., Gleitman, H., & Gleitman, L. R. (1991). On the semantic content of subcategorization frames. *Cognitive psychology*, *23*(3), 331-392.

Ford, M., Bresnan, J., & Kaplan, R. M. (1982). A competence based theory of syntactic closure. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp. 727–796). Cambridge, MA: MIT Press.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 748-775.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37, 58–93.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language processing*.

Psychology Press.
Gibson, E., Schutze, C. T., & Salomon, A. D. A.-J. (1996). The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research*, *25*(1), 59–92
Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
Goldberg, A. E . (2004). "Pragmatics and Argument Structure." *The handbook of pragmatics*. L. Horn, & G. Ward (eds.). John Wiley & Sons.
Goldberg, A. E. (2011). Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics*, *22*(1), 131-154.
Goldberg, A. E. (2013). Substantive learning bias or familiarity effect? Comment on Culbertson, Legendre and Smolensky (2012). Cognition 127 (3) 420-426
Goldberg, A. E., Casenhiser, D., & Sethuraman, N. (2004). Learning Argument Structure Generalizations. *Cognitive Linguistics, 15*, 289-316.
Green, G. (1974). *Semantics and syntactic regularity*. Bloomington: Indiana University Press.
Gries, S. T., & Divjak, D. (Eds.). (2012). *Frequency effects in language learning and processing* (Vol. 1). Walter de Gruyter.
Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.
Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The Learnability and Acquisition of the Dative Alternation in English. *Language, 65*(2), 203–257.
Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press.
Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford University Press.
Hawkins, J. A. (2014). *Cross-linguistic Variation and Efficiency*. Oxford University Press.
Hovav Rappaport, M., & Levin, B. (2008). The English dative alternation: The case for verb sensitivity. *Journal of Linguistics*, *44*(01), 129–167.
Hudson Kam, C. & E. Newport (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* 1, 151 –195.
Jaeger, F. T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23-62.
Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588-611.
Lakoff, G. (1970). *Irregularity in Syntax*. New York: Holt, Rinehart and Winston.
Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. University of Chicago Press.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.
Langacker, R. (1987). *Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites*. Stanford University Press.
Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.
Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic

reduction. *Advances in neural information processing systems*, *19*, 849.
Lewis, J. D., & Elman, J. L. (2000). Learnability and the Statistical Structure of Language: Poverty of Stimulus Arguments Revisited. In *Annual Boston University Conference on Language Development*. Boston University.
Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208.
MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101, 676–703.
Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, 126, 313-318.
Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2005). The role of frequency in the acquisition of English word order. *Cognitive Development*, *20*(1), 121–136.
Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, *26*(4), 393–424.
Oehrle, R. (1976). *The Grammatical Status of the English Dative Alternation*. PhD Thesis, MIT. Cambridge: The MIT Press.
Payne, J., Pullum, G. K., Scholz, B. C., & Berlage, E. (2013). Anaphoric *one* and its implications. *Language*, *89*(4), 794–829.
Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam: Benjamins.
Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(03), 607-642.
Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122, 280-291.
Pierce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2):8-13.
Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, Mass: MIT Press/Bradford Books.
Pierrehumbert, J. (2001). Exemplar dynamics: word frequency, lenition and contrast. *Typological Studies in Language*, *45*, 137–158.
Poser, William J. 1992. Blocking of phrasal constructions by lexical items. Lexical matters, ed. by Ivan Sag and Anna Szabolsci, 111–130. Stanford: CSLI.
Reali, F., & Christansen, M. H. (2005). Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. *Cognitive Science*, *29*(6), 1007–1028.
Rendell, L. (1986). A general framework for induction and a study of selective induction. *Machine Learning*, *1*(2), 177-226.
Robenalt, C. & Goldberg, A. E. (to appear). Judgment and frequency evidence for Statistical Preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics.*
Saffran, J. R. (2001). Words in a sea of sounds: the output of infant statistical learning. *Cognition*, *81*, 149–169.

Saffran, J. R. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

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

Schumacher, R. A., Pierrehumbert, J. B., & LaShell, P. (2014). Reconciling Inconsistency in Encoded Morphological Distinctions in an Artificial Language Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci2014) Austin , TX : Cognitive Science Society.

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

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(33), 11629–34.

Suttle, L., & Goldberg, A. E. (2011). The partial productivity of constructions as induction. *Linguistics*, *49*(6), 1237-1269.

Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind*. Oxford University Press.

Thompson, S. A. (1990). Information Flow and Dative Shift in English Discourse. In J. Edmondson, K. Feagin, & P. Mühlhäusler (Eds.), *Development and Diversity: Linguistic Variation across Time and Space.* (pp. 239–253). Summer Institute of Linguistics.

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

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

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

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

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

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

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