The contextual modulation of semantic information

Aaron Steven White
Johns Hopkins University

Valentine Hacquard
University of Maryland

Philip Resnik
University of Maryland

Jeffrey Lidz
University of Maryland

We investigate how the semantic information that is carried by a word's linguistic contexts is modulated by the kind of nonlinguistic context that word is used in—focusing in particular on verbs. Using a corpus analysis of child-directed and child-ambient speech, we first show that the distribution of syntactic structures and lexical items that a verb occurs with is more complex in some contexts (meal contexts) than in others (play contexts). This finding raises the question of whether meal contexts have more semantic information than play contexts. We address this question in a large-scale Human Simulation Paradigm experiment, where we show that the relationship between distributional complexity and semantic information is not direct. Rather, nonlinguistic contexts with less distributional complexity (e.g. play contexts) yield better learning when participants (simulated learners) have little information, and nonlinguistic contexts with greater distributional complexity (e.g. meal contexts) yield better learning when those learners have more information. Further, simulated learners are only able to take advantage of the information in the higher-complexity contexts if they have access to both syntactic and lexical information, while in contrast, those learners are able to take advantage of the information in the lower-complexity contexts even if they have access to only syntactic information. Based on this, we suggest that linguistic cues from some contexts (e.g. play contexts) may be used to form initial imprecise hypotheses about a verb's semantics, while linguistic cues from other contexts (e.g. meal contexts) may be used to refine those initial hypotheses.

Keywords: semantic information, distributional complexity, entropy, context, language acquisition, syntactic bootstrapping, lexical bootstrapping

1 Introduction

Successful word learning relies heavily on learners' ability to extract semantic information about a word from that word's contexts of utterance. But semantic information is not distributed equally across such contexts (Medina, Snedeker, Trueswell, and Gleitman, 2011; Trueswell, Medina, Hafri, and Gleitman, 2013). For example, contexts in which the utterer of a concrete noun is pointing or gazing at a potential referent for (a phrase containing) that noun tend to be more semantically informative than ones without such pointing or gazing (Baldwin 1991,9; Bloom 2000; Bruner 1983; Frank, Goodman, and Tenenbaum 2009; Nappa, Wessel, McEldoon, Gleitman, and Trueswell 2009; Tomasello 1992; Yurovsky and Frank 2015; among many others).

Contexts of utterance themselves fall into *context classes* that are likely to be important for determining how semantic information is distributed (Yont, Snow, and Vernon-Feagans, 2003). A trivial case of this might be that, when children are learning words for foods, utterances of those words in, e.g., a kitchen or dining room are likely to carry more semantic information as a consequence of the higher likelihood that the food will be present (and thus a potential object of gaze or ostension). In this hypothetical case, we might say that the kitchen contexts are more informative about a particular class of words (food words) than other contexts.

Beyond ostension and joint attention, many features of a context of utterance can contribute to the semantic information that is extractable from that context. For instance, the distribution of linguistic elements surrounding a word—e.g. the lexical items a word occurs near (Gillette, Gleitman, Gleitman, and Lederer, 1999; Snedeker, 2000; Snedeker, Gleitman, and Brent, 1999), the syntactic structures a word occurs in (Brown, 1957,7; Fisher, Gleitman, and Gleitman, 1991; Gillette et al., 1999; Gleitman, 1990; Landau and Gleitman, 1985; Lederer, Gleitman, and Gleitman, 1995; Naigles, Gleitman, and Gleitman, 1993; Naigles, 1990,9), and the joint distribution of the two (Arunachalam and Waxman, 2010; Yuan and Fisher, 2009, cf. Resnik 1996)—is known to be useful as a cue to that word's meaning.

But is the semantic information coming from linguistic cues also modulated by context class? And if so, how? Understanding this modulation is important for understanding word learning specifically, not least because learning at least certain classes of words—e.g. verbs—is likely heavily reliant on linguistic cues (Gleitman, 1990; Gleitman, Cassidy, Nappa, Papafragou, and Trueswell, 2005). But it is also important for understanding learning more generally, since it contributes to our knowledge of how different distributional structures found in learners' environments are related to the information that can be extracted from that environment.

Preliminary evidence for the modulation of semantic information from linguistic cues comes from the fact that adult-directed speech is more distributionally complex than child-directed speech on various measures (Buttery, 2006; Buttery and Korhonen, 2005). In particular, the variety of syntactic structures that particular verbs occur in is higher in adult-directed speech than in child-directed speech (cf. Roland and Jurafsky 1998, 2002 who discuss genre-based effects in adult-directed language).

But though this finding is suggestive, it does not directly indicate whether semantic information coming from linguistic cues is also modulated by context class. Higher distributional complexity in a particular context class may provide more evidence about words' syntactic and semantic affordances, which is known to be useful for word learning (Bunger and Lidz 2004; Naigles 1996, cf. Rowe 2013 and references therein), but even mature language speakers' ability to use such evidence as a basis for inference is likely to be limited by memory and processing demands. Indeed, even when only considering very basic object-word pairings as evidence for word meaning (Yu and Smith, 2007,1), adults show inability to use all (or even most) of the available information (Medina et al., 2011; Trueswell et al., 2013), suggesting that we should not equate distributional complexity that an ideal learner could employ with distributional complexity that even mature speakers of a language can employ.

Further, though there is indirect evidence that context class modulates semantic information (cf. Yont et al., 2003), relatively little is known about how it does so. And much of what is known is based on comparisons of the specific syntactic structures and lexical items that parents use during toy play, book-reading, and mealtime, not necessarily properties of their distribution. For instance, Hoff-Ginsberg (1991) shows that the syntactic structures that middle-class mothers employ during book-reading tend to be more complex than those they employ during toy play. Weizman and Snow (2001) show that this extends to low-income mothers, and Salo, Rowe, Leech, and Cabrera (2015) show that it extends to low-income fathers. Leech and Rowe (2014) show that book genre—chapter v. picture—also modulates complexity, with picture book-reading tending to elicit a higher quantity of more complex speech.

One reason book-reading tends to involve speech that is more syntactically complex is that it involves more metalinguistic talk than in toy play (Jones and Adamson, 1987). Metalinguistic talk in turn tends to involve heavy use of clausal embedding, since many words used to talk about linguistic acts are clause-embedding verbs—e.g., *say*, *tell*—thus driving up the syntactic complexity of particular sentences (Ely, Gleason, MacGibbon, and Zaretsky, 2001)—though not necessarily the distributional complexity of the input as a whole. Ely et al. show that such metalinguistic talk is also highly prevalent in meal contexts (see also Blum-Kulka, 1997; Masur and Gleason, 1980). In contrast to book-reading, where metalinguistic talk tends to be about the book at hand, the large amount of metalinguistic talk in meal contexts involves reports of past speech (Ely and Gleason, 1995), some of which arises as a consequence of adult-to-adult speech that children are privy to.

One possibility that this raises is that meal contexts involve a mix of child- and adult-directed (but *child-ambient*) speech that may in turn give rise to a modulation in distributional complexity similar to that found by Buttery and Korhonen (2005). If this is true, contrasting meal contexts, on the one hand, with book-reading and toy play contexts, on the other, might also prove to be useful for understanding how the semantic information carried by linguistic cues is related to distributional complexity, and it would speak to the general learning question mentioned above: how do different distributional structures found in learners' environments (their input) relate to the information that can be extracted from that environment (their intake)?

We take this investigation up in this paper. We focus in particular on verbs' syntactic and lexical distributions, since as we noted above, learning even the simplest verbs—e.g. action verbs like *run* and *kick*—likely requires linguistic distributional information (Gleitman, 1990); and learning more semantically complex verbs—e.g. mental state verbs like *think* and *want*—from only nonlinguistic situational cues is likely to be a nonstarter (Gillette et al., 1999; Gleitman et al., 2005; Papafragou, Cassidy, and Gleitman, 2007).

We begin in Section 2 by comparing the distributional complexity of child-directed speech in toy play/book-reading contexts to the distributional complexity of child-directed speech in meal contexts—in particular, dinner contexts. We show that, even controlling for variability due to particular speakers and particular verbs, verbs' distributions are more complex, on average, in dinner contexts than they are in toy play/book-reading contexts.

This raises the question of whether the higher distributional complexity found in dinner contexts translates into better learning from distributions found in those contexts compared to toy play and book-reading contexts. To address this question, we employ a modified version of the Human Simulation Paradigm (HSP Gillette et al., 1999) that allows us to explicitly measure how close the semantic representation a participant learns over the course of training is to the true verbs' semantic representation.

In Section 3, we review methodologies in the HSP family, focusing in particular on recent approaches that use the

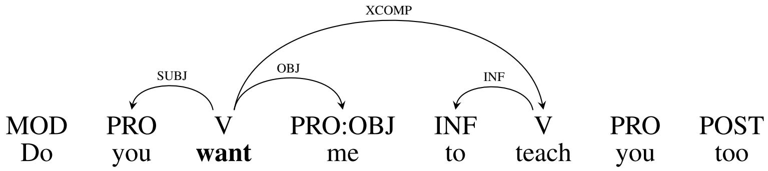

Figure 1. Example of dependency parse arrows relevant to subcategorization frame extraction for *want*. Example from the Gleason corpus (dinner transcript for Bobby) from CHILDES.

one-shot HSP to norm standard HSP studies (Medina et al., 2011; Trueswell et al., 2013). In Section 4, we present our own one-shot HSP norming study focused on measuring the semantic information in linguistic contexts drawn from play and meal contexts. We also present a similarity judgment-based extension of the one-shot HSP that allows us to assign partial credit to inaccurate responses. We show that, as measured by accuracy (the standard measure in HSP), there is no detectable difference, on average, between play and meal contexts in terms of the semantic information carried by particular linguistic contexts, but as measured by similarity, individual play contexts have more semantic information, on average, than dinner contexts.

In Section 5, we use the results of our one-shot HSP study to construct our modified version of HSP, the Spatial Human Simulation Paradigm (SHSP), which uses similarity judgments instead of categorical responses to measure semantic information. We show that play/book-reading contexts yield better learning when adult participants—henceforth, *simulated learners*—have fewer learning instances and that meal contexts yield better learning when simulated learners have more learning instances. Further, simulated learners are only able to take advantage of the information in the meal contexts if they have access to both syntactic and lexical information, while they are able to take advantage of the information in the play/book-reading contexts even if they have access to only syntactic information.

In Section 6, we conclude by discussing how these results relate to higher-level issues of input and intake in language acquisition (Lidz and Gagliardi, 2015; Omaki and Lidz, 2015). In particular, we suggest that context class may play an important role in the development of children's semantic representations for verbs: semantic information extracted from some context classes may be used to form initial imprecise hypotheses about a verb's semantics while semantic information extracted from other context classes may be used to refine those initial hypotheses.

2 Measuring distributional complexity

In this section, we establish that the classes of contexts of utterance that children find themselves in differ with respect to distributional complexity. In particular, we show that *meal* contexts—specifically, *dinner* contexts—contain higher amounts of distributional complexity with respect to verbs—both in terms of syntactic distributional complexity and in terms of lexical distributional information—than *play* contexts. To operationalize this claim, we use the information-theoretic *(self-)entropy* of a distribution (Shannon, 1948) as a measure of distributional complexity. This measure is useful because it allows us to simultaneously quantify the variety of syntactic structures and lexical items that a particular verb occurs with as well as the probability of seeing that verb with any one of those structures or lexical items. All analysis code is available on the first author's github.

2.1 Corpus

To compare syntactic and lexical distributional complexity between dinner and play context, we use the Gleason (1980) corpus, obtained from CHILDES (MacWhinney, 2014a,1). A description of this corpus, taken from the CHILDES manual for North American English corpora (p. 44), is as follows.

> The participants are 24 children aged 2;1 to 5;2 who were recorded in interactions (a) with their mother, (b) with their father, and (c) at the dinner table. The 24 participants were recruited through nursery schools and similar networks, and were from middle-class families in the greater Boston area. There were 12 boys and 12 girls. All families were White, and English was spoken as a first language in all families. Each child was seen three times: once in the laboratory with the mother; once in the laboratory with the father; and once at dinner with both mother and father. The laboratory sessions were videotaped and audiotaped, and the dinners were only audiotaped. Laboratory sessions included: (a) play with a toy auto, (b) reading a picture book, and (c) playing store.

Only 22 of the 24 children (11 females) have transcripts for both the dinner session and the play session, and so all of the following analyses are based on this subset of 22—the same 22 that Ely et al. (2001) include in their analysis.

2.2 Data extraction

To extract each verb's syntactic and lexical distributions, we used the morphosyntactic annotations that ship with the Gleason corpus, which were constructed using the MOR and POST morphological analyzers (Parisse, 2000) and the MEGRASP dependency parser (Sagae, Davis, Lavie, MacWhinney, and Wintner, 2007). For each child, we first concatenated the morphosyntactically annotated transcripts of the mother and father laboratory sessions, which we henceforth refer to as the play transcripts. Then, for each of the play and dinner transcripts, we extracted utterances made by either the mother or the father—excluding the participating child, the experimenter, and any other children (relevant only for the dinner session).

For each of the utterances extracted in this fashion, we extract the set of verbs that occur in the utterance using the part-of-speech (POS) annotations available in the MEGRASP dependency parses. This verb was then stripped of any inflectional morphology using the MOR/POST annotations. For instance, the verb *thinks* would be morphologically analyzed as *think-3s*, and so we would strip off *3s* to yield *think*.

To obtain the lexical distribution for each verb extracted in this fashion, we extract every word in the utterance besides the verb and strip those words of any inflectional morphology. For instance, *dogs* would be morphologically analyzed as *dog-PL*, and so we would strip off *PL* to yield *dog*. We do not remove stop words like *the* or *of*, since these may be important features of a verb's linguistic distribution that we do not capture in the syntactic structures we extract. We then count the number of times that a particular verb showed up with a particular word across utterances in a context class for a particular child and divide by the total number of words that that verb occurred with in that context class for that child. This yields the maximum likelihood estimate for the distribution over words, keyed to a particular verb for a particular child in a particular context class.

To obtain the syntactic distribution for each verb, we use the MEGRASP dependency parse to construct subcategorization frames. The procedure is fully documented in the analysis code that we have released, but roughly, our extractor records all dependencies coming out of the verb, and in the case of clausal complements, it also records some dependencies coming out of the embedded verb (in order to determine tense and complementizer information for the embedded clause). Figure 1 shows an example of the pieces of a MEGRASP dependency parse relevant to determining the syntactic structure associated with *want*. The frame we extract from this parse is *NP _ NP S[-tense]*. We then count the number of times that a particular verb showed up with a particular subcategorization frame in a context class for a particular child and divide by the total number of times that that verb occurred in that context class for that child. This yields the maximum likelihood estimate for the distribution over subcategorization frames, keyed to a particular verb for a particular child in a particular context class.

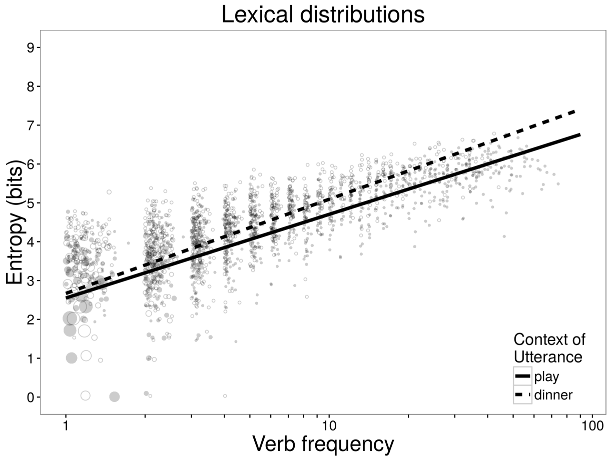

Figure 2. Entropy of lexical distributions for each verb by child and context class

2.3 Measure

For both the lexical and the syntactic distributions, we calculate the self-entropy for each verb-child-context class tuple. As noted above, self-entropy is useful for our purposes because it simultaneously quantifies the variety of syntactic structures and lexical items that a particular verb occurs with as well as the probability of seeing that verb with any one of those structures or lexical items. For instance, the formula for the entropy H of the maximum likelihood estimate of the subcategorization frame distribution $\hat{p}$ for a particular verb-child-context class tuple is

$$H(\hat{p}) = -\sum_{\text{frame} \in \text{subcats}} \hat{p}(\text{frame}) \log \hat{p}(\text{frame})$$

where terms in the sum such that $\hat{p}(\text{frame}) = 0$ are set to 0. Thus, entropy is highest when every lexical item or subcategorization frame that a verb occurs with is equally likely—the distribution is uniform—and it is lowest when only a few frames or lexical items are very likely. Entropy also captures differences in lexical item and subcategorization frame variety. This can be seen most straightforwardly for uniform distributions. Because a uniform distribution over n objects always has an entropy of $\log n$, such a uniform distribution will always have higher entropy than a uniform distribution over $n - 1$ objects, since $\log n > \log n - 1$.

2.4 Results

We now compare the distribution of lexical item and subcategorization frame entropy in dinner and play contexts using linear mixed effects models. We consider three predictors of entropy: CONTEXT CLASS (*play*, *dinner*), LOG VERB

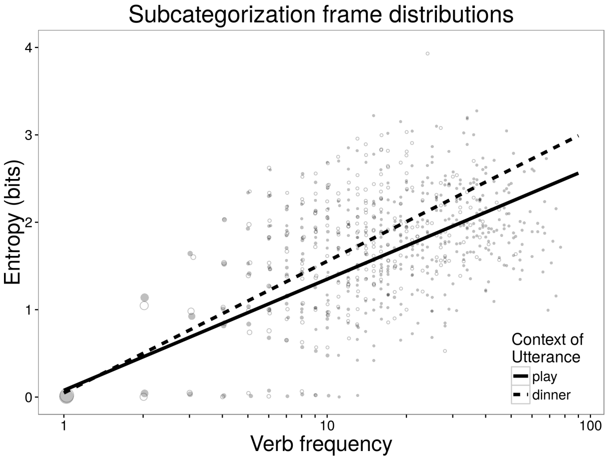

Figure 3. Entropy of subcategorization frame distributions for each verb by child and context class

FREQUENCY, and their interaction. Verb frequency is an important variable to control for because, if a verb has a true distribution that is highly entropic but the verb is only seen a few times, our estimate of its distribution could have much lower entropy than the true distribution. This is in part because if a verb's frequency is lower than the total number of distinct subcategorization frames or lexical items, the maximum possible entropy for the estimate of a verb's distribution is the log of that frequency (see discussion in the previous section). More precisely, if there are K distinct subcategorization frames or lexical items, the maximum possible entropy for a verb seen N times is $\log \min(N, K)$.

Beyond fixed effects for CONTEXT CLASS (*play*, *dinner*), LOG VERB FREQUENCY, and their interaction, we also include the maximal fixed effects structure for these data: random intercepts for child and verb as well as by-child and by-verb random slopes for CONTEXT CLASS (*play*, *dinner*), LOG VERB FREQUENCY, and their interaction. Thus, we allow the model to explain variability in (i) the overall complexity of each child's parents' speech, (ii) how this complexity changes as a function of CONTEXT CLASS, LOG VERB FREQUENCY, and their interaction, (iii) the overall complexity of each verb's distribution, and (iv) how this complexity changes as a function of CONTEXT CLASS, LOG VERB FREQUENCY (across children), and their interaction.

Our hypothesis is that, even after controlling for the four kinds of variability modeled by the random effects structure, *dinner* contexts will show higher entropy than *play* contexts in at least syntactic distributions, but possibly also lexical distributions. Assuming a dummy coding with *play* as the reference level for CONTEXT CLASS, this might arise in two ways: (i) as a simple effect of CONTEXT CLASS, which would mean that, across the frequency spectrum, the *dinner* context would show a positive effect on entropy; or (ii) as an interaction between CONTEXT CLASS and LOG VERB FREQUENCY. This second scenario is more likely in light of the fact that lower frequencies have a ceiling equal to the log frequency, and thus seeing an effect on the lower end of the frequency spectrum will be harder. Indeed, for verbs that are only seen once, the entropy is necessarily zero for the subcategorization frame entropy and thus, for subcategorization frames, the only effect we expect is a positive interaction between *dinner* and LOG VERB FREQUENCY. Thus, the crucial effect we must see to confirm our hypothesis is the interaction.

We begin by fitting the full model with lexical item entropy as the dependent variable. We then test for the presence of an interaction between CONTEXT CLASS and LOG VERB FREQUENCY by removing it from the fixed effects (but retaining the relevant by-child and by-verb random slopes) and refitting the model. We find that, according to a log-likelihood ratio test, the interaction is significant ($\chi^2(1) = 6.04$), $p < 0.05$). As can be seen in in Figure 2, the interaction is furthermore positive, and because the reference level is *play*, this means that the average difference in entropy between between the *dinner* and *play* contexts grows positively with verb frequency.

Because entropy is always positive, one potential worry here is that we may not have appropriately modeled skewness and heteroscedasticity in the conditional distribution of lexical entropy. A White test suggests that there is heteroscedasticity relative to LOG VERB FREQUENCY ($\chi^2(5) = 7.60$), $p < 0.01$). This is not particularly surprising, since as we noted above, lower frequency verbs will tend to have lower entropy as a function of how well we were able to estimate their true distribution. This may not be problematic for the estimates themselves as long as the conditional distribution is itself symmetric. We cannot test for symmetry directly, but we do note that the mean (0.00) and median (0.05) are extremely close, suggesting little skew. Further, the correlation between residual entropy LOG VERB FREQUENCY is extremely small (0.00, 95% CI=[-0.03, 0.03]), suggesting that skewness does not arise only in particular regions of the frequency spectrum.

Next, we attempt to fit the full model with subcategorization frame entropy as the dependent variable. This model does not converge due to the optimizer's inability to a singular Hessian, which in turn appears to be due to a high correlation (very near -1) between the CONTEXT CLASS random slopes and the random intercepts for both children and verbs. We thus refit the model without random slopes for the interaction between CONTEXT CLASS and LOG VERB FREQUENCY but retaining the fixed effects interaction for CONTEXT CLASS and LOG VERB FREQUENCY. We then test for the presence of an interaction between CONTEXT CLASS and LOG VERB FREQUENCY by removing it from the fixed effects and refitting the model. We again find that, according to a log-likelihood ratio test, the interaction is significant ($\chi^2(1) = 67.78$), $p < 0.001$). As can be seen in Figure 3, the interaction is furthermore positive, and because the reference level is *play*, this means that the average difference in entropy between the *dinner* and *play* contexts grows positively with verb frequency.

We again check for skewness and heteroscedasticity in the conditional distribution of the subcategorization frame entropy. A White test suggests that, as before, there is heteroscedasticity relative to LOG VERB FREQUENCY ($\chi^2(5) = 9.11$, $p < 0.001$). Again, this is not particularly surprising, since as we noted above, lower frequency verbs will tend to have lower entropy as a function of how well we were able to estimate their true distribution. This may not be problematic for the estimates themselves as long as the conditional distribution is itself symmetric. We cannot test for symmetry directly, but we again note that the mean (0.00) and median (-0.04) are extremely close, suggesting little skew. Further, the correlation between residual entropy LOG VERB FREQUENCY is extremely small (0.00, 95% CI=[-0.03, 0.03]), suggesting that skewness does not arise only in particular regions of the frequency spectrum.

2.5 Discussion

Using entropy as a measure of complexity, we have established that the contexts of utterance that children find themselves in differ with respect to distributional complexity. In particular, we showed that *meal* contexts—specifically, *dinner* contexts—contain higher amounts of distributional complexity with respect to verbs—both in terms of syntactic distributional complexity and in terms of lexical distributional information—than *play* contexts. This is consistent with previous research on toy play, book-reading, and dinner contexts that found differences in the complexity of the linguistic forms used: in general, parents use less complex linguistic structures in toy play compared to both book-reading and meal contexts. We have added to this research by showing that beyond modulation of complexity in particular linguistic structures, we also find modulation of complexity in the distribution of these forms.

This finding raises the question of whether the higher distributional complexity found in dinner contexts translates into better learning from distributions found in those contexts compared to toy play and book-reading contexts. To answer this question, we use a modified form of the Human Simulation Paradigm and focus in on verbs that tend to occur in syntactically complex structures—specifically, verbs that take subordinate clauses. Given the findings in the abovementioned prior work, if the increased distributional complexity found in dinner contexts carries increased semantic information about any sort of word, it seems likely to be verbs that take subordinate clauses. There is also preliminary evidence from work in the Human Simulation Paradigm, reviewed below, that this may be the case.

3 Measuring semantic information

In this section, we review prior work that uses the Human Simulation Paradigm, which is a standard suite of instruments used in word-learning research to measure the semantic information that some feature of a context of utterance carries about a word (Gillette et al., 1999; Snedeker, 2000; Snedeker et al., 1999). We first describe the original paradigm, which has been used to measure the semantic information carried by both nonlinguistic and linguistic contextual features. We then review a recent modification to this paradigm employed by Medina et al. (2011) and Trueswell et al. (2013) to explicitly manipulate the amount of semantic information available to participants, focusing on nonlinguistic contextual features. In Section 5, we employ insights from both the classic paradigm and Medina et al.'s modification to construct our study, which focuses on linguistic contextual features.

3.1 The classic paradigm

In the standard task, adult participants are given some information about the context a word was uttered in—scenes, surrounding words, structures, etc.—and they are asked which word occurred in that context. The accuracy with which they can recover the true word given the context (or a set thereof) is then used as a measure of informativity.

The idea behind using adult learners here is that, while there is a question of what level of conceptual development child learners have attained at a particular age, adult learners presumably have some level of conceptual development that allows them to grasp the meanings of the sorts of words children learn in the first few years of life. This in turn allows one to ask questions about the informational properties of various purported learning cues, controlling for conceptual development.

In Gillette et al. 1999, six different contextual conditions are tested, which form the basis for later work in HSP. Their first experiment investigates the usefulness of solely nonlinguistic scene information plus lexical category information. For this experiment, they chose the 24 most frequent nouns and the 24 most frequent verbs from a transcript of video-recorded play sessions. They then collected 6 video clips for each verb and played participants each of these clips in sequence, with a beep occurring when the word occurred. Participants were asked to guess at each beep which word occurred in that position, having been told (i) that the beeps for a particular set of clips all involved the same word and (ii) whether the word was a noun or verb. Participants did much better with nouns than with verbs for this experiment. They further showed that this could be wholly predicted by the imageability rating for particular nouns or verbs, where verbs like *run* are rated much more imageable than propositional attitude verbs like *think*.

Gillette et al.'s second task, which is the one important for current purposes, manipulated the kind of information participants had access to for making inferences about the word meaning: (i) scene information only—the original task—(ii) items from lexical categories (noun, verb, adjective) that surround the word; (iii) scene information plus lexical information; (iv) syntactic frames with nonce words replacing words from lexical categories; (v) the full sentence the word occurred in (syntactic frame + word from lexical categories); and (vi) all of the previous kinds of information. The upshot is roughly that more information is better.

Focusing in on the verbs, Snedeker and Gleitman (2004) replicate this more-information-is-better result (see also discussion in Gleitman et al., 2005). They argue that verbs fall into three groups with respect to the sorts of cues used to learn them (and further that these three groups correspond to well-defined development stages): "relatively concrete verbs that describe specific actions in and on the observable world (*fall*, *stand*, *turn*, *play*, *wait*, *hammer*, *push*, *throw*, *pop*), the more abstract mental-content verbs (*know*, *like*, *see*, *say*, *think*, *love*, *look*, *want*), and a third set, of what have been called light verbs (*come*, *do*, *get*, *go*, *have*, *make*, *put*)." Relevant for current purposes, the mental-content verbs (propositional attitude verbs) have low accuracies in the contexts (i-iii) from above but show a spike in accuracy when moving to contexts (iv-vi)—i.e. when one has syntactic information, possibly in concert with lexical information.

Papafragou et al. (2007) replicate this strong effect of syntactic information on ability to infer propositional attitude verb meanings, but they also show that scene information is not totally irrelevant. If participants are given scenes involving something that highlights a character's beliefs, participants are more willing to conjecture propositional attitude verb meanings. This is further reinforced by combining these scenes with linguistic—specifically, syntactic—information.

3.2 Norming HSP with HSP

A question that has arisen recently as a topic of debate in the word-learning is how memory constrains learners' ability to utilize cues to a word's meaning (Medina et al., 2011; Trueswell et al., 2013). In many accounts of noun-learning in particular, the assumption is often that learners have access to the full history of their experience with a word, or at least some nonnegligible chunk (see Yu and Smith 2012 for recent discussion). One way memory constraints have been investigated is to ask how sensitive to the information carried by a particular piece of contextual information a learner's hypotheses are. The idea is that, to the extent that learners decisions are based on particular learning instances—the less smooth their learning trajectory—the more constrained their memory for previous instances is likely to be. Testing this requires some way of measuring the information carried by particular learning instances.

Following Gillette et al. (1999), Medina et al. (2011) selected the 24 most frequent nouns and 24 most frequent verbs from a video corpus developed by one of the authors. They then selected 288 from this corpus and presented each of 37 participants with 96 scene-only vignettes (2 per word). Unlike the standard HSP, in which participants would see 6 clips in a row for the same word, each of the clips in the Medina et al. study was completely disconnected from the others. We henceforth refer to this method as *one-shot HSP*. The idea here is that, by disconnecting the vignettes, the informativity of each can be measured on its own terms. They take as their measure of informativity, the accuracy with which participants can recover the actual word that occurred in the vignette (at the point of a beep, as is standard).

Medina et al. threshold items into two sets—High Informativity (HI) and Low Informativity (LI)—by their accuracy, choosing 33%, as this gave them a 1:5 ratio of HI:LI vignettes. They then use this partitioning to construct sequences of five vignettes with one HI instance and four LI instances, thus using the experiment with disconnected vignettes as a norming experiment for the second, a standard HSP task. They find that manipulating the placement of the HI instances significantly affects participants' ability to eventually recover the correct word. In particular, the earlier a high informativity instance the better.

No verb or abstract noun vignettes show up as HI vignettes, so Medina et al. drop verbs and abstract nouns from consideration for the rest of the experiments. This is important for us, since we focus only on such verbs for the purpose of our experiment. In the next section, we describe two norming tasks with similar aims to the one employed by Medina et al., but adapted for use with syntactic frames instead of scene information.

4 Norming tasks

We now present two norming tasks that are used to construct the materials for our large-scale HSP task. As in Medina et al.'s norming tasks, our aim is to measure the semantic information that isolated syntactic and lexical contexts carry about a word.

The first task is a one-shot HSP task similar to the one employed by Medina et al., but using linguistic context instead of scenes. In analyzing the results of this task, we follow prior work in HSP in using accuracy as a measure of informativity and find that the distribution of semantic information in dinner and play contexts is not significantly different, though participants' ability to recover the true syntactic category (verb) is.

While accuracy is useful as a rough measure of semantic information, accuracy has known issues as such a measure (see discussion in Gillette et al., 1999), since it does not take into account responses that may be semantically close to the true word but which are nonetheless inaccurate. Thus, no distinction is made between bad guesses, such as ones that don't even fall into the correct syntactic category, and better guesses, which might involve words that are semantically close to the true word. This second problem is exacerbated when an item might be very informative about the semantics of the word that occurred in it, but when a semantically related high frequency word—plausibly, one that comes to mind more quickly—also fits well within the context.

We remedy this in the second task, where we use semantic similarity judgments to measure how close in meaning participants' responses in the first task were to the true word. This allows us to assign partial credit to incorrect responses. We find that, when such partial credit is assigned, play contexts have slightly higher amounts of semantic information than dinner contexts, on average.

4.1 One-shot HSP norming task

4.1.1 Design. The first norming task takes the form of a one-shot HSP task with only linguistic contexts. In this task, participants are given a sentence with a blank somewhere in it and are asked to fill the blank with the word they think most people would respond with. All sentences were sampled from a corpus of child-directed speech as described below, and thus the blank replaces a real word. For instance, (1-a) was an actual sentence used in the experiment.

(1) a. I **told** you I'm not having a new baby now.
b. I florped you I'm not having a new baby now.

This task was conducted online, and responses were collected using an HTML text box. This text box was filled with a greyed out placeholder verb *florp* that disappeared when a participant started typing, which we implemented using the standard HTML `input` tag `placeholder` attribute. This placeholder verb had tense/aspect morphology matching the verb that it takes the place of. For instance, (1-b) shows the sentence derived from the true sentence (1-a). Text boxes were autofocused to allow participants to use only the keyboard while performing the task.

The norming items fell into one of 40 conditions as formed from a full cross of the following three factors whose descriptions are given in subsequent sections: VERB (levels: 10 attitudes verbs), LEXICAL CONTENT (levels: *real*, *nonce*), and CONTEXT CLASS (levels: *dinner*, *play*). Participants received one item from each condition for a total of 40 items. (2) and (3) give examples of nonce and real items from both the dinner and play contexts for *say*.

(2) *dinner* sentences
a. *real:* You had your chance to have soup and you said you didn't want anymore.
b. *nonce:* You had your zeep to gloll wooth and you said you didn't slonch anymore.

(3) *play* sentences
a. *real:* What is he saying to the lady?
b. *nonce:* What is he saying to the florn?

In addition to providing typed responses to the HSP task, participants were asked a memory question about the item they just responded to in order to ensure they were paying attention. The rationale and construction of this memory task is described further below.

4.1.2 Materials. For each of the 31 verbs investigated in White, Hacquard, and Lidz under review, all sentences containing at least one of those verbs were extracted from the Gleason corpus. White et al.'s set of verbs was used as a basis for material construction (i) because they were chosen to span the semantic classes of clause-embedding verbs that are common in the semantics literature and (ii) because White et al. collected similarity judgments for all of these verbs, which are publicly available on White's github and which we employ for our large-scale study.

In this first norming task, as for the corpus analysis reported above, both the *dinner* and *play* contexts were considered as separate factor levels of CONTEXT CLASS. The number of sentences each verb occurred in within both the play and dinner sections of the corpus (across children) were then tabulated, and the top ten most frequent verbs calculated. For each of these high frequency verbs, up to 30 sentences were sampled from the dinner sessions and up to 60 sampled from the play sessions (30 from the mothers' play sessions and 30 from the fathers').

Each of these sentences was then hand-checked for transcription errors and acceptability. Unacceptable sentences were marked for exclusion, including those that might be unacceptable or hard to parse out of context. This is necessary given the nature of the one-shot HSP task, in which participants do not receive any surrounding linguistic context. All such modifications to the sample are recorded in the online materials released on the first author's github.

After this acceptability checking and modification procedure was complete, 20 sentences were subsampled for each verb from each modified context set (play and dinner), excluding the unacceptable sentences. These sentences form the *real* level of the LEXICAL CONTEXT factor.

To create the items in the *nonce* level of the LEXICAL CONTEXT factor, all nouns, verbs, adjectives, and adverbs for the above sentences were replaced with nonce words with morphology matching the ones found on the real words found in the original sentence. All determiners, prepositions/particles (of, to, at, up, etc.), and complementizers (that, if, for, to) were retained. Among the determiners were included quantificational determiners/quantifiers (*every(thing)*, *any(thing)*, etc.) and WH words (*who*, *what*, *where*, etc.). The intention here was to retain only words from functional categories.[1] To this end, some nouns and verb-like elements were also retained.

These noun exceptions included all personal pronouns (*I*, *you*, *(s)he*, *me*, *mine*, etc.) as well as temporal (*now*, *then*) and locative indexicals (*here*, *there*).[2] These exceptions seem reasonable since under many theories, they fall into the determiner class or are at least partially constituted by a determiner-like meanings.

Verb exceptions included all auxiliary verbs: all forms of *be*, perfect auxiliary forms of *have*, all modal auxiliaries (*can*, *might*, *must*, etc.). Semi-modals (*have to*, *ought to*) were treated as lexical verbs in this respect—i.e. *have* or *ought* would be replaced with a nonce word but *to* retained.

Finally, four common adverbs were excepted from replacement by a nonce word: *too*, *either*, and *else*. This seems reasonable since, in contrast to derived adverbs like *carefully* or *intentionally*, these adverbs' meanings are logical in nature and thus naturally fall into a class with, e.g., the pronouns. Indeed, all are anaphoric.

4.1.2.1 Memory task. Of the 20 real word items participants received, half were followed by the question "which word was in the previous sentence?" along with five words, only one of which was actually in the previous sentence. (None of the nonce word sentences were followed by this memory task.) For instance, participants who saw the sentence in (4) received the memory question along with the set of words in (5).

(4) I think what we should do is try to florp what we took apart last and put that together first.

(5) a. Which word was in the previous sentence?
b. {boy, mothers, together, never, family}

Participants' response accuracy for each item was then collected and analyzed for the purposes of data validation.

4.1.3 Participants. Participants were recruited until each item had at least 20 observations associated with it after the data validation procedure described in the last section. Participants were allowed to respond to up to three lists. 577 unique participants were recruited through Amazon Mechanical Turk (AMT) using a standard Human Intelligence Task (HIT) template designed for externally hosted experiments and modified for the specific task. Of these unique participants, 483 responded to a single list, 88 responded to two lists, and 6 responded to three lists. No participant that responded to multiple lists responded to the same list twice.

Prior to viewing the HIT, participants were required to score seven or better on a nine question qualification test assessing whether they were a native speaker of American English. Along with this qualification test, participants' IP addresses were required to be associated with a location within the United States, and their HIT acceptance rates were required to be 95% or better. After finishing the experiment, participants received a 15-digit hex code, which they were instructed to enter into the HIT. Once this submission was received, participants were paid for their time.

4.1.4 Data validation. Three data validation techniques were used. First, for each participant, the number of correct responses to the memory task were tabulated (by list for participants that responded to more than one list). The vast majority of participants (0.793) obtain perfect scores, with almost all of the remainder answering incorrectly only once (0.168) or twice (0.031). Given this distribution, only participants that scored 8 (of 10) or better on the memory task were retained. This resulted in the exclusion of 3 participants: 1 who responded with only 3 correct, 1 that responded with only 6 correct, and 1 that responded with only 7 correct. (Even if the participant with 7 correct were retained at this stage of validation, most of that participant's responses would be excluded in the third stage due to the fact that that participant gave almost solely nonce word responses.)

Next, participants' log reaction times (log RTs) were analyzed. First, each participant's median log RT and the interquartile range(IQR)—the difference between 25^{th} and 75^{th} precentiles)—of their log RTs were computed. The median and IQR of each of these statistics was then computed over participants. Participants were excluded using Tukey's method, wherein the Tukey interval ([Q1-1.5*IQR, Q3+1.5*IQR]) is constructed for both by-participant medians and IQRs and participants excluded if their median log RT or IQR log RT fell outside this interval. No participant's median log RT fell outside the Tukey interval of median log RT over participants and no participant's IQR log RT fell outside the Tukey interval of IQR log RT over participants; thus no participants were excluded under these criteria.

The median log RT-based exclusion procedure was also conducted for particular responses. For each participant, the IQR of the log RT for that participant's responses was computed. Responses were then excluded if they fell below the participant-specific Tukey interval. 6 responses (across participants) were excluded in this way.

[1] There is a question here whether all prepositions are purely functional. This seems unlikely, but the replacement of prepositions can severely degrade participants ability to access the syntactic structure of a sentence. This is likely due to the fact that prepositions are relatively closed-class—at least compared to nouns, verbs, and adjectives. It is standard in human simulation paradigm experiments to retain prepositions (cf. Gillette et al., 1999).

[2] Temporal expressions like *today*, *yesterday*, and *tomorrow* might fall under this criterion, since they seem indexical in ways similar to *now* and *then*. The problem is that many complex temporal expressions, like *last night* or *next week* are similarly indexical, and it is unclear where to draw the line. One criterion could be to retain only single word indexical expressions, like *yesterday*, *today*, and *tomorrow*, but this privileges expressions that involve days over those that involve other time intervals, and thus some amount of arbitrariness is necessary. The current methods seems to us the most conservative if indexical expressions are to be retained.

The final filtering step was to exclude all nonword responses. As an approximation, nonword was defined as any word not occurring at least once in the PukWaC corpus (Baroni, Bernardini, Ferraresi, and Zanchetta, 2009). Of the 26953 total response tokens and 1517 response types, 258 response tokens and 85 response types were marked as nonwords in this way. Those words that were marked as nonwords were then handchecked. Many cases either involved the participant responding with the placeholder verb—i.e. *florp*, *florps*, *florping*, or *florped*—a random string[3]—e.g. *lmpw* or *toxat*—or a multiword string[4]—e.g. *where were you* or *he asked me*.

Other responses, however, were clear typos. (Indeed, multiple participants emailed to apologize for having made a typo somewhere in the experiment.) For instance, *know* had three typo variants—*knlw*, *knkow* and *knokw*. When their correct variant was clear, these typos were corrected manually. For instance, *knkow* would be changed to *know*.[5] These corrections results in 16 of the datapoints originally marked as nonword responses to become word responses. The remaining 242 were then excluded.

One problem with using corpus counts to filter nonwords is that, while filtered words will tend to be nonwords (the method has high precision) some nonfiltered words may still be nonwords, since common typos will be counted (the method has lower recall). A subsequent filtering step was thus conducted by hand. Of the 26711 response tokens and 1432 remaining after the first nonword filtering step, 75 response tokens from 62 response types were deemed nonwords that were not typo variants of a real word.[6] Of the 26636 response tokens and 1372 response levels remaining after this filtration, 44 were clear typo variants of true words, which were corrected. The final number of response tokens after filtering was thus 25636 and the final number of response types after filtering was 1328.

4.1.5 Results. We conduct three analyses in this section. In the first, we show that LEXICAL CONTEXT but not CONTEXT CLASS has a significant effect on accuracy. In the second analysis, we assess the extent to which participants' inaccurate responses are due to an inability to recover the correct syntactic category of a word. This is important, since if participants cannot recover the syntactic category, other higher order properties of the linguistic context, such as the syntactic structure, probably won't be accessible either. In the final analysis, we give a qualitative characterization of how close participants' inaccurate guesses were in general, which we use to motivate our second task.

4.1.5.1 Accuracy. Figure 4 plots the distribution of accuracy computed by item for each pairing of CONTEXT CLASS and LEXICAL CONTEXT. The accuracy for each item (1 of 20 from its particular condition) is calculated as the number of times the true verb—e.g. *told* in (6-a)—was given as a response to the item created from that sentences—e.g. (6-b)—divided by the number of responses to that item.

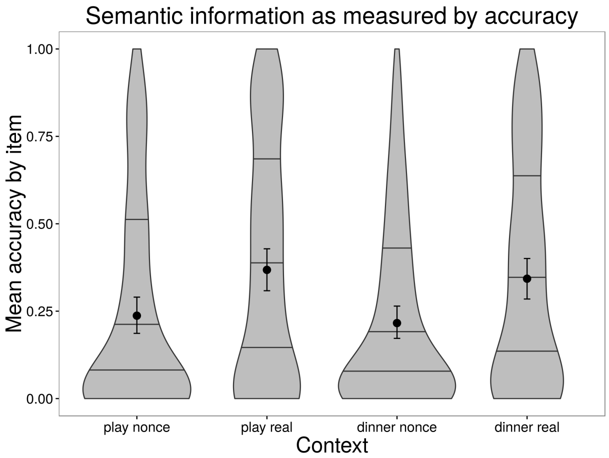

Figure 4. Distribution of accuracy computed by item. Horizontal lines show 1^{st}, 2^{nd}, and 3^{rd} quantiles; points show mean accuracy; and error bars show 99% confidence intervals for those means computed by nonparametric bootstrap.

(6) a. I **told** you I'm not having a new baby now.
b. I florped you I'm not having a new baby now.

Horizontal lines show first, second, and third quantiles; points show mean accuracy; and error bars show 99% confidence intervals for those means computed by a nonparametric bootstrap.

We see that, on the whole, sentences whose content words are replaced with nonce words have lower accuracy. This is not surprising given that these sentences are designed to remove some information that might help participants infer the meaning of the word in the blank. Further, the CONTEXT CLASS that a context of utterance was found in does not seem to affect accuracy.

This pattern is corroborated by the results of fitting a mixed effects logistic regression to these accuracy data with accuracy as the dependent variable, fixed effects of LEXICAL CONTEXT, CONTEXT CLASS, and their interaction, and random intercepts for participants, verbs, and items

[3] Some participants' strategy in this case was to type a nonword from the sentence itself. For instance, two of the nonwords used in the experiment were *spurply* and *slargle*, and these were both given as responses.

[4] Participants were explicitly instructed not to do this at multiple points in the instructions.

[5] These changes were not made to the raw data itself, but rather in the analysis script, and are documented in the analysis scripts made available on the first author's github.

[6] In fact, for some of these nonwords, the intent was clear. For instance, *practic* is likely a typo of *practice*. These typos were only corrected if the true variant already showed up at least once elsewhere (not as a typo).

(nested under verbs). We first test the interaction between LEXICAL CONTEXT and CONTEXT CLASS by refitting the model without this interaction but retaining the random effects structure. This test did not reach significance ($\chi^2(1) = 0.13, p = 0.719$), and thus the interaction term was dropped. The same procedure was carried out for LEXICAL CONTEXT ($\chi^2(1) = 64.08, p < 0.001$) and CONTEXT CLASS ($\chi^2(1) = 0.11, p = 0.741$). Only LEXICAL CONTEXT was significant under this criterion. Thus, LEXICAL CONTEXT but not CONTEXT CLASS appears to modulate semantic information, as measured by accuracy.

We now look to the variance estimates for the random effects to try to understand what gives rise to the modulation of semantic information that we do see. The variance for the participant random intercepts is 0.173 (sd: 0.416) (in log-odds space); the variance for the verb random intercepts is much larger at 3.183 (sd: 1.784); and the variance for the item random intercepts was similarly large at 2.946 (sd: 1.716). This means that participants varied little in their ability to answer accurately—for comparison, the participant random intercept standard deviation is less than half the size of the estimated fixed effect of *lexical context*. Verbs and items on the other hand show much higher variability. For instance, the verbs *tell*, *know*, *think*, and *want* show higher accuracy than other verbs, like *remember*, *guess*, and *hear*.

One possible explanation for the verb variability may be a frequency effect: verbs with higher frequencies may come to mind more readily during the task and so participants might also respond with these more readily. But if participants are more willing to respond with higher frequency verbs, those verbs might have a higher accuracy due to their frequency. To investigate this, a second mixed model was fit with TRUE WORD LOG FREQUENCY—obtained from the counts available from the ukWaC (Baroni et al., 2009) website—as a predictor alongside LEXICAL CONTEXT. (The same random effects structure was retained.) TRUE WORD LOG FREQUENCY is significant as per a log-likelihood ratio test comparing this new model against model with only LEXICAL CONTEXT ($\chi^2(1) = 6.82$, $p < 0.01$). A third model was fit with the interaction between TRUE WORD LOG FREQUENCY and LEXICAL CONTEXT, but this term was not significant LEXICAL CONTEXT ($\chi^2(1) = 0.04$, $p = 0.85$).

Investigating the regression coefficients, we find that the effect of TRUE WORD LOG FREQUENCY (0.93) is of approximately the same size as that of LEXICAL CONTEXT (1.10). This means that for every order of magnitude increase in frequency of the true word, participants were more accurate (on average) to about the same extent as if they had gotten a *real* item instead of a *nonce* item. Surprisingly, however, controlling for frequency does not affect the variance estimates for verb random intercepts much, though there is a reduction in this estimate. (One would not expect it to affect the participant or item random intercepts, since frequency is a property of verbs, not participants or items, and indeed, those estimates remain constant.) The new estimate of the variance for verb intercepts is 2.156 (sd: 1.469), which is about an 18% reduction in the standard deviation from the previous estimate. This suggests that some, but by no means all, the variability in verb responses is driven by frequency effects.

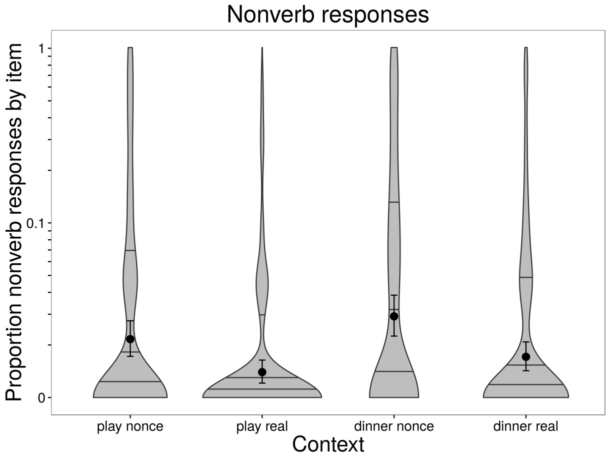

Figure 5. Distribution of nonverb response proportions computed by item. Horizontal lines show 1^{st}, 2^{nd}, and 3^{rd} quantiles; points show mean accuracy; and error bars show 99% confidence intervals for those means computed by nonparametric bootstrap.

What else might drive accuracy? To get an accurate response, it must be clear from a particular item which verb fits in that item. On the one hand, as we show below, many of the most common responses to particular items are verbs that, while not accurate because they don't match the true item, nonetheless share a semantic component with the true item and are thus intuitively closer in meaning than some random item. This would suggest an item that is somewhat informative about the semantics of a word, but not fully informative. On the other hand, some items may be extremely uninformative—to the point that even the syntactic category of the true response is unclear. These two states of affairs are quite different in nature, and ideally there would be some way of pulling them apart.

Next, we investigate how easy it is to recover the correct syntactic category (verb) across items. We then turn to a preliminary analysis of which syntactic features are most useful in giving an accurate response. And finally, we give a qualitative analysis of the inaccurate responses that participants give in order to assess how informative items are about the meaning a verb has.

4.1.5.2 Nonverb responses. Going back to at least Brown 1957, it has been known that syntactic category is a useful cue to word meaning. Indeed, to use the syntactic distribution a verb occurs in to help infer its meaning, one first needs to know that the word they are dealing with is a verb in the first place. One possibility for at least some of the inaccuracy for some of the above verbs then could be uncertainty about the syntactic category that the word falls into.

To assess the overall uncertainty about syntactic category, responses were labeled for whether they were a verb or not by hand. Figure 5 shows the distribution nonverb response proportions computed by item. A model with the same structure as the one discussed in the last section was fit with NONVERB as the dependent variable, and the likelihood ratio test procedure repeated. In this case, as before, the interaction between LEXICAL CONTEXT and CONTEXT CLASS is not significant ($\chi^2(1) = 0.04$, $p = 0.85$); but in contrast to the previous case, both main effects of LEXICAL CONTEXT ($\chi^2(1) = 15.74$, $p = 0.001$) and CONTEXT CLASS ($\chi^2(1) = 5.14$, $p < 0.05$) are significant.

The effect of CONTEXT CLASS is a positive effect of *dinner* contexts, meaning that participants give more nonverb responses for *dinner* items. This effect is interesting, since in the accuracy model, no significant effect of CONTEXT CLASS is found. This means that though participants are not reliably more likely to respond with the true word based alone on the context in which a sentence was uttered, they do more reliably infer that the word is a verb in the *play* contexts than in the *dinner* contexts. In both cases, however, the probability of nonverb responses is quite low, which can be seen by the fact that we must use a log scale in Figure 5 to visualize the response distributions.

Dinner sentences were longer on average, and so one plausible driver of the increase in nonword responses is that, in longer sentences, participants get overloaded with the number of nonce words whose category they are uncertain about and then shut down, making a random guess.[7] To test this, a model including item word count alongside the other predictors as well as one including both item word count and its interaction with LEXICAL CONTEXT alongside CONTEXT CLASS was constructed. Neither model shows significant improvement in likelihood ratio tests ($\chi^2(2) = 1.546, p = 0.462$). Thus, it does not appear to be length that gives rise to higher amounts of nonverb responses (and lower accuracy).

Another possibility is that certain aspects of a sentence's syntactic structure are easier or harder to parse when the lexical items are nonce. This could give rise to higher accuracy in certain cases—likely also dependent on the verb and its subcategorization frame distribution—and lower accuracy in others—even to the point that participants cannot recognize that the word they are trying to guess is a verb.

To assess how the syntax of a sentence affects nonverb responses, we use a standard method for measuring variable importance: mean gini decrease in a random forest (Breiman, 2001). The idea behind this method is to build many small decision trees that attempt to predict whether a participant gave a nonverb responses based on the verb at hand, the LEXICAL CONTEXT, the CONTEXT CLASS, and various syntactic features of the sentences.

First, the syntactic features of every item were hand-coded using the feature set similar to the one used in our analysis of distributional complexity: COMPLEMENTIZER (*none*, *finite*, *polar question*, *WH question*), EMBEDDED TENSE (*none/no embedding*, *infinitival*, *bare*, *gerund*, *tensed*), MATRIX SUBJECT (*none*, *referential*, *it*, *there*), EMBEDDED SUBJECT (*none/no embedding*, *nominative*, *accusative*, *case unknown*), MATRIX OBJECT 1 (*true*, *false*), MATRIX OBJECT 2 (*true*, *false*), MATRIX OBLIQUE (*true*, *false*). These features were then merged with their corresponding items in the HSP dataset and, along with LEXICAL CONTEXT and CONTEXT CLASS, were entered into a random forest classifier with 1000 trees and three variables tried at each split.

We find that a single syntactic feature—MATRIX SUBJECT—is found to be more important than LEXICAL CONTEXT as a predictor of nonverb responses with the remainder of the syntactic features being of negligible importance. In investigating this variable, we find that imperatives (7) and subjectless questions (8) drive its importance. Specifically, when the verb does not have an overt subject, participants gave more nonverb responses.

(7) **florp = say**
  a. Florp you are recording your dinner for David.
  b. Florp you are wheecing your veigh for David.

(8) **blarp = think**
  a. Blarp you will play with him tomorrow?
  b. Blarp you will yock with him swen?

This is interesting from the point of view of previous findings that the distribution of imperatives in parental speech is correlated with children's vocabulary; specifically, more imperatives in a child's input predicts a smaller vocabulary (Yont et al., 2003). The current finding suggests a potential causal factor underlying this correlation: in certain contexts, imperatives may make it harder to determiner an unknown word's syntactic category.

4.1.5.3 Response distributions. Another source of inaccuracy might be participants giving inaccurate-yet-semantically-similar responses. This is not taken into ac-

[7] To establish that dinner sentences were longer on average, a mixed effects poisson regression was conducted on only the LEXICAL CONTEXT:*real* sentences with sentence length as the dependent variable, a fixed effect for CONTEXT CLASS, and random intercepts for VERB. The CONTEXT CLASS effect is significant under a likelihood ratio test comparing this model to one without the fixed effect ($\chi^2(1) = 5.08, p < 0.05$). The coefficient for the *dinner* level is furthermore positive (log increase in length: 0.084), suggesting that dinner sentences are longer on average. (Only LEXICAL CONTEXT:*real* were used in this regression since the LEXICAL CONTEXT:*nonce* sentences will necessarily have the same length as their corresponding LEXICAL CONTEXT:*real* sentence.)

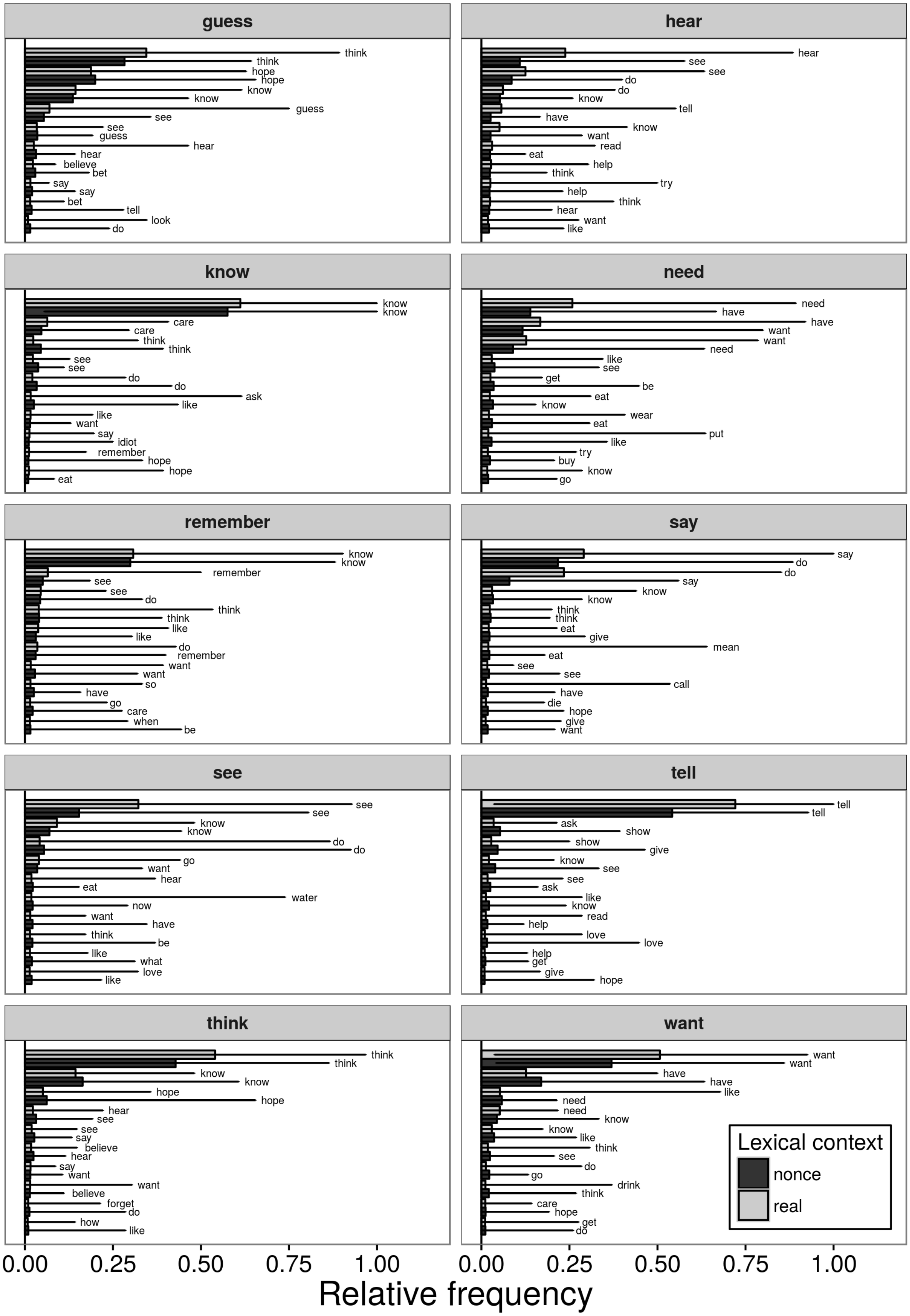

Figure 6. Distribution of responses for each verb. Bars give the median relative frequency over items; errors bars give the range of those relative frequencies over items.

count in the previous analyses. However, even if participants don't give the exact verb that occurred in a particular position, they might nevertheless answer with one that is semantically similar. For instance, in our analysis of the random effects components of our accuracy model, we noted that the overall accuracy for *remember* is lower than average. To some extent, this could be a product of nonverb responses—e.g. in analyzing the random intercepts from the nonverb models, it turns out that the intercept for *remember* is reliably positive, suggesting that part of the explanation for the low accuracy of *remember* is that people gave more nonverb responses.

But even among verb responses, the sorts of responses participants give are far from random. For instance, the most common response to *remember* sentences was not *remember*, but *know*. Considering that *know* seems to be involved in the meaning of *remember* at some level of representation—remembering something presupposes knowing something—it's quite interesting that participants give this response.

To get a feel for how prevalent inaccurate-yet-semantically-similar responses are, we now turn to a qualitative analysis of the distribution of responses to each verb's items, which we augment with a quantitative analysis in the next section. To delve into these responses, it will be useful to first find the proportion of times a word was given for a particular item, then look at the distribution of those proportions. This is analogous to looking at the distribution of accuracy over items shown in Figure 4, where the outcomes are binary (accurate v. nonaccurate); the differences is that, here, the outcomes are treated as many-valued and plotted by verb.

Prior to looking at these distributions, it is necessary to perform some preprocessing on the responses. Because some items will involve an inflected version of the verb, we first map the true verb and the responses to their root form. This *lemmatization* was done by hand for all response types, regardless of syntactic category (except for adverbs).

After this lemmatization, the proportion of times a particular root form was given as a response to a particular item was computed. This results in a relative frequency distribution over the root form of responses for each item. Because (i) for current purposes we care about general trends over items for particular verbs and (ii) it is difficult to visualize each item's distribution in an easy to digest way, we graph these distributions by response type in Figure 6. This graph shows, for each verb, the response roots with the highest median relative frequency over items. Thus, as in 4, each bar represents datapoints for 20 items in the relevant condition. The bar itself gives the median and the error bar gives the range.

We see that for verbs *think*, *want*, *see*, *tell*, and *know*, the most common responses in both the *real* and *nonce* conditions are the true verb itself. This suggests that many of the sentences these verbs are found in are highly informative about their semantics and, further, that this seems to be a result of their syntax to a great extent—since these verbs are guessed even without lexical context. Among the other responses for these verbs, *think* and *know*, notably, have responses that tend to involve beliefs: the second most common response to *think* sentences is *know* in both *real* and *nonce* conditions and the third most common response to *know* sentences is *think*. *Tell* has responses that tend to be communicative (*ask* and *show*) or that at least involve a transfer (*give*). *See* has responses that are cognitive but nonperceptual (*know*) but also some very bleached responses (*do*). And *want* has a secondary response that is quite bleached (*have*) but other responses that are closer to its semantics, which involves preferences (*need*)—though these responses are not given much more often than belief responses like *know* and *think*.

For the verbs *say*, *need*, and *hear*, the most common response in the *real* conditions is the true verb, but this is not true in the *nonce* conditions. This may suggest that much of participants' ability to guess the correct verb in these three cases is somehow dependent on the lexical context. Each has a slightly different response profile in terms of how (intuitively) close the responses are to the true verb's meaning. For instance, like *want*, *need* receives many bleached responses (*have*), but it also receives quite a few responses with a desire meaning (*want*) in both the *real* and *nonce* conditions. Similarly, *hear* receives quite a few bleached response (*do*), but also quite a few perceptual responses (*see*). This may suggest that the fact that *hear* is perceptual is encoded in at least some of the items—e.g. in small clause items, like *John heard Mary leave*—but that the kind of perceptual verb it is is hard to ascertain.[8] *Say* differs from *need* and *hear* in this respect in the sense that, while it has common belief-type responses (*think* and *know*), it doesn't seem to have any common speech responses, except for maybe *call*. One possibility is that *say*—like many of the speech verbs—has better perceptual correlates, and thus that a syntax-based learner would require more nonlinguistic context than we provided.

The final group of verbs—*remember* and *guess* are interesting for the fact that, though their most common responses in either the *nonce* or *real* conditions are technically incorrect, they are both quite close semantically. In the case of *remember*, the vast majority of responses are *know*. This is interesting since *remember* seems to encode *know* as a subpart of its meaning (as mentioned above). Further, the other responses to *remember* are also broadly involve beliefs, with both cognitive (*think*) and perceptual (*see*) verbs. *Guess*

[8] The fact that *see* is a common response to both *see* and *hear* sentences could suggest a frequency effect, which seems likely since *see* is about two orders of magnitude more frequent (as measured by the frequency in ukWaC; Baroni et al. 2009) in the *present* tense than *hear*. (They are about equally frequent in the past.)

similarly has a most common response that involves belief (*think*), and many of the other responses also involve beliefs (*know*, *see*).[9]

4.2 Similarity norming task

We noted that despite the wide variability in item accuracy across verbs, even inaccurate responses are not completely random. Indeed, this has been noted since the inception of HSP. In this task, we aim to quantify this semantic closeness using a similarity judgment task built from the responses recording in the previous task. We then conduct analyses of these similarities akin to the accuracy analyses from the last section.

4.2.1 Design. All response roots that were (i) marked as verbs in the previous task and (ii) were inaccurate were paired with the true verb that occurred in that position. (This is why only inaccurate pairs were retained. An accurate pair is just the same verb twice.) There were 2429 such pairs. For each of the 10 verbs sampled from the corpus, the pairs involving that word as the true word were then randomized and inserted into lists, with amount of pairs proportional to the number of unique response types to a particular word. With the criterion that each list should contain around 60 pairs, 37 lists were created in this way.

4.2.2 Participants. 155 participants were recruited through Amazon Mechanical Turk (AMT)—five for 36 of the lists and 10 for the last[10]—using a standard Human Intelligence Task (HIT) template designed for this particular experiment. All qualification requirements were the same as for the first norming task.

4.2.3 Data validation. As in the previous task, a log reaction time-based data validation procedure was conducted. First, each participant's median log RT was computed. The median of these median log RTs as well as the interquartile range (IQR)—the difference between 25^{th} and 75^{th} precentiles)—was then computed. Participants were excluded using Tukey's method applied to both participant medians (described above) and IQRs. 5 participants' median log RTs fell below Q1 log RT minus 1.5 times the IQR and were thus excluded. No participant's IQR fell outside of the Tukey interval of IQRs across participants.

The same RT-based exclusion procedure was also conducted for particular responses. For each participant, the IQR of the log RT for that participant's responses was computed. Responses were then excluded if they fell below that participant's median log RT minus 1.5 times that participant's specific IQR. 18 responses (across participants) were excluded in this way. This yielded a total of 12320 observations with the minimum number of observations per item being 4. (Post-filtering, 7 items had 4 responses and the remaining 1737 had 5 or more.)

4.2.4 Results. Prior to analysis, z-scoring was applied to participants' ratings. The average of the z-score transformed variants of the judgments was then taken, and each result was associated with each of the true word-response pairs from the previous experiment. Figure 7 shows the distribution of these by-item similarity means by context of utterance, excluding both accurate responses and nonverb responses.

To assess the effects of LEXICAL CONTEXT and CONTEXT CLASS on the similarity between an inaccurate verb response and the true word, a linear mixed effects model was fit with mean z-score similarity for each true word-response word as the dependent variable; LEXICAL CONTEXT, CONTEXT CLASS, their interaction, and TRUE WORD LOG FREQUENCY as fixed effects; and random intercepts for participant, verb, and item (nested under verb). Thus, this model has the same structure as the original accuracy model, but instead of being fit to accuracy as the dependent variable, it is fit to similarity as indexed by the mean of participants' z-scored similarity responses to a particular response-true word pair.[11] And as before, likelihood ratio tests were conducted to assess the significance of particular predictors. As before, the interaction term was not significant ($\chi^2(1) = 0.01, p = 0.92$) and was thus dropped. The two main effect terms LEXICAL CONTEXT ($\chi^2(1) = 15.12, p < 0.001$) and CONTEXT CLASS ($\chi^2(1) = 7.14, p < 0.01$) were significant and were thus kept.

The effect of CONTEXT CLASS is a small positive effect of the *play* context class that is about 75% of the size of the LEXICAL CONTEXT effect and about 80% of the size of the TRUE WORD LOG FREQUENCY effect. This effect is interesting, since in the accuracy model, no significant effect of CONTEXT CLASS is found. This means that though participants are not reliably more likely to respond with the true word based alone on the context in which a sentence was uttered, they do get reliably closer to that word. This may in

[9] One interesting aspect of both of these cases is that the belief verbs cross not only the cognitive-perceptual divide, but also the factive-nonfactive divide. Factive *remember* gets *think* responses and nonfactive *guess* gets factive responses *know*. Indeed, this extends to even the previous two groups of verbs, where nonfactive *think*, *say*, and *hear* got *know* responses and factive *know* got *think* responses. One reason this may be is that the purported syntactic cue to factivity—that the factive verb occurs with both polar question and nonquestion complements (see White et al. under review for discussion)—cannot be contained within a single subcategorization frame; it is fundamentally an aspect of a verb's distribution.

[10] The five extra participants were recruited because an off-by-one error that affected only one list was discovered after the first five were run for that list.

[11] Indeed, the current model can be thought of as the continuous component of a two-stage hurdle model: one that first considers whether the response given by a participant will be a verb or not; then decides whether that response will be accurate; then if inaccurate, decides how similar the response is to the true response. The accuracy and nonverb models from the last section would serve as the first two components.

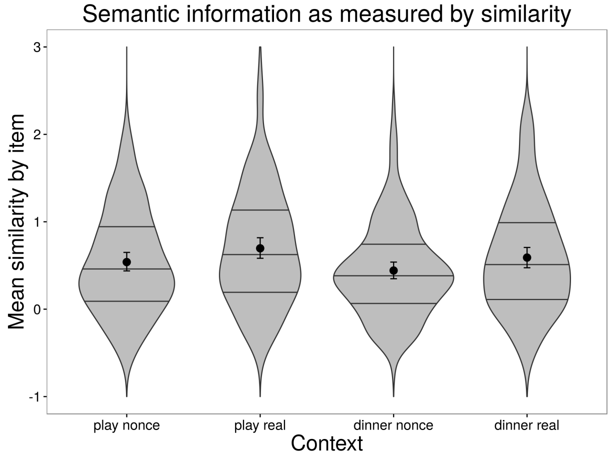

Figure 7. Distribution of mean *z*-scored similarity computed by item. Horizontal lines show 1^{st}, 2^{nd}, and 3^{rd} quantiles; points show mean accuracy; and error bars show 99% confidence intervals for those means computed by nonparametric bootstrap.

turn arise due to the same reason that participants are better able to grasp the true syntactic category of the word in the *play* contexts. It further suggests that *dinner* contexts may be a particularly interesting test case, since on the whole they provide somewhat less information per-occurrence about an attitude verb's semantics.

4.3 Discussion

We presented two norming tasks aimed at gathering a measure of similarity of participants' responses to the true verb that occurred in a particular item. We found that, as measured by accuracy, there is no evidence for a difference in the semantic information conveyed by isolated items, comparing play and dinner contexts; but there is a difference in how easily participants can infer the correct syntactic category and how semantically close their inaccurate guesses are. By these latter two measures, play contexts are more informative on average.

In conjunction with our finding that dinner contexts show higher distributional complexity, this section's finding raises the question of whether the semantic information conveyed by multiple items together is higher in the dinner or play contexts. On the one hand, it could be that the higher average informativity of particular items in the play contexts compounds to yield higher semantic information when participants receive multiple items. On the other hand, it could be that the informativity of particular items in the play context shows diminishing returns, with the dinner context items becoming more informative when considered jointly.

5 Spatial human simulation

In this section, we use the two norming tasks presented above to construct an experiment aimed at investigating whether we see contextual modulation of semantic information when considering multiple items. This paradigm is close to a standard HSP experiment in the sense that, unlike the first norming task above, participants are told that they are learning the same word over multiple items. It differs, however, in the sense that, instead of giving a free choice response after each item is presented, participants are asked for similarity judgments after the entire set of items.

We manipulate both the number of training items that participants receive as well as the relative informativity of those items as measured by accuracy and similarity from the norming tasks. Our main finding is that nonlinguistic contexts with less distributional complexity (e.g. play contexts) yield better learning when simulated learners have little information, and nonlinguistic contexts with greater distributional complexity (e.g. meal contexts) yield better learning when learners have more information. Further, simulated learners are only able to take advantage of the information in the higher-complexity contexts if they have access to both syntactic and lexical information; in contrast, those learners are able to take advantage of the information in the lower-complexity contexts even if they have access to only syntactic information.

5.1 Design

The task has two main parts: a training phase, in which participants receive a set of sentences containing the same novel word, and a test phase, in which participants are asked to make ordinal scale similarity judgments. In the first part of the test phases (*test phase 1*), participants make similarity judgments between the novel word they were just trained on and all 31 real words from the ordinal scale experiment conducted by White et al. (see above).

In the second part of the test phase (*test phase 2*), participants make similarity judgments between two known words drawn from this same group and selected so as to span the similarity range. This second part is used to assess how close a particular participant's real word-to-real word judgments are to the average found by White et al.. We then use this measure to control for possible differences in participants' semantic representations.

The experiment has four two-level factors: CONTEXT CLASS (*dinner* v. *play*), LEXICAL CONTEXT (*real* v. *nonce*), INFORMATIVITY (*high* v. *low*) and TRAINING SIZE (*big* v. *small*). These latter two factors are explained in more detail below; the former two are the same as from the norming studies. For each condition defined by these four factors we created materials for each of the 10 verbs tested in the norming studies.

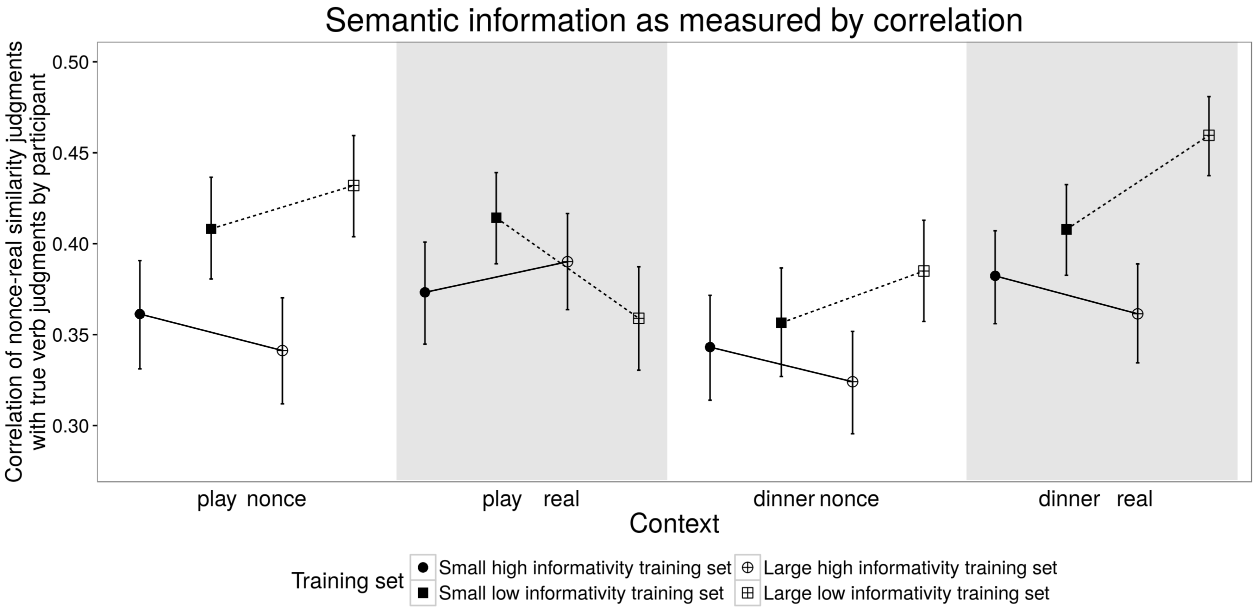

Figure 8. Relationship between context class and participants' performance in test phase 1 (nonce-real similarity judgments).

5.2 Materials

Training sets were constructed by partitioning the sentences corresponding to each verb from the previous experiment into two sets by-verb. For each of the ten verbs from the norming studies, the median ridit scored similarity value was obtained by averaging over the values for responses to each item, including accurate responses as 1 and nonverb responses as 0.[12] Items with scores below this median were labeled low informativity (LI) for that verb and items with scores above the median were labeled high informativity (HI). Thus, for each verb at each level of LEXICAL CONTEXT (*real* and *nonce*), there were 10 LI and 10 HI items.

Training sets were then constructed from either solely HI or solely LI items. This diverges from Medina et al. 2011 in the sense that their training sets involved a mix of the two sets. We use the current method in order to provide optimal conditions for seeing how the semantic information coming from particular utterances compounds to yield more than the sum of their parts. That is, it allows us to ask, once we have controlled for the informativity of particular utterances, do we gain anything from the distribution itself, over and above what we predict from the particular utterances?

Two of the training sets use all of the items in the informativity partition (TRAINING SIZE: *big*). The other uses only half the items (TRAINING SIZE: *small*). In the LI + small case, the lowest informativity items were used—i.e. those in the first quartile of informativity scores for their particular verb—and in the HI + small case, the highest informativity items were used—i.e. this in the fourth quartile of the informativity scores for their particular verb.

Ten different nonce-real test sets were constructed. This test set consisted of 31 pairs: the nonce verb participants were trained on paired with the 31 verbs in the ordinal scale task used to construct the norming task (White et al., under review). A real-real test list that remained constant across training sets was also constructed. This list was selected from all pairs in the aforementioned ordinal scale task by ordering those pairs based on their mean z-scored rating across participants and then taking every 30^{th} pair. This selection was hand-checked to ensure that a few verbs didn't show up a disproportionate amount of times under this procedure. None did. The reasoning behind this selection procedure was to ensure that the pairs come from across the similarity space. As mentioned above, this was done so that any participant specific aspects of the similarity responses could be controlled for.

5.3 Participants

4800 participants (710 unique) were recruited through Amazon Mechanical Turk (AMT) using a Human Intelligence Task (HIT) template designed for this particular experiment. All qualifications requirements were the same as those used for the norming tasks.

[12] Ridit scoring was used here because it is unclear how to combine accurate and nonverb responses with the similarity judgments when normalizing the similarity judgments using z-scoring. Further, White et al. (under review) show that the assumptions underlying ridit scoring are likely more accurate relative to how participants actually make similarity judgments. Indeed, we use z-scoring above only because it is the *de facto* standard normalization method.

Participants were allowed to do as many of the lists as they liked, though they were not allowed to do the same list more than once. Lists were deployed in batches of 10, each containing a training set for a particular true verb. This was done to ensure that any participant who did two lists in quick succession would not have gotten two lists pertaining to the same true verb. The median number of lists that each participant did was 2 and the mean was 6.8.

5.4 Data validation

Three separate filtering stages were conducted prior to analysis: (i) participant filtering based on memory task accuracy; (ii) participant filtering based on median and IQR of (log) reaction times to the similarity task; and (iii) response filtering based on median and IQR of (log) reaction times by-participant.

For the first stage of filtering a mixed effects logistic regression with random intercepts for participant and verb was built with accuracy on the memory task as the dependent variable and all experimental conditions (LEXICAL CONTEXT, ITEM INFORMATIVITY, and TRAINING SIZE) as well as all possible two- and three-way interactions as fixed effects. The inclusion of these fixed effects was meant to control for the fact that the memory task in certain conditions may be harder—e.g. those conditions with LEXICAL CONTEXT: *nonce*, where participants had to remember nonce words—or less well estimated—e.g. an error in the conditions with TRAINING SIZE:*small* counts more than one in TRAINING SIZE:*big*. This full interaction model was tested against a model without the three-way interaction but with the two-way interactions using a likelihood ratio test, and the three-way interaction was found to be significant ($\chi^2(1) = 20.12$, $p < 0.001$), so the full model was kept.

Participants were excluded based on the Best Linear Unbiased Predictors (BLUPs) of the participant random intercepts inferred by the model. These BLUPs for the participant intercepts were then mean-centered and standardized by their standard deviation. All participants whose standardized intercept fell below -2—i.e. two standard deviations below the mean accuracy—were then excluded. This results in the exclusion of 35 total participants, and the loss of 21514 total datapoints.

Next, as in previous sections, participants were excluded based on a reaction time analysis (see above for procedure). Only reaction time to the similarity judgment task was considered. In the median-based filtering, 10 participants had a median log RT below the Tukey interval of participant medians and were excluded, resulting in the exclusion of 14083 datapoints. In the IQR-based filtering, 17 participants had IQRs of log RTs outside the Tukey interval, resulting in the exclusion of 8091 datapoints.

Finally, as in previous sections, particular responses were excluded based on a reaction time (see above for procedure).

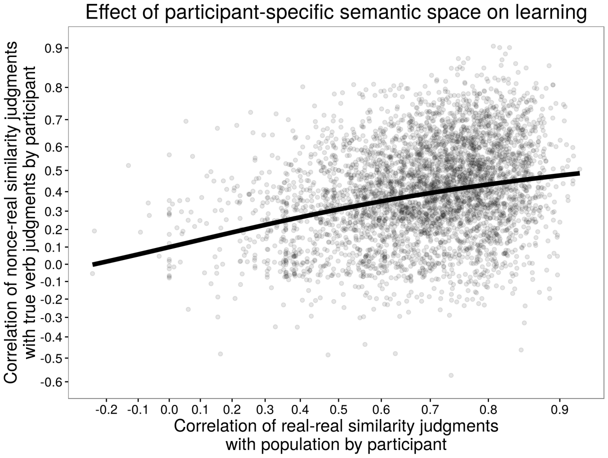

Figure 9. Relationship between participants' performance in test phase 2 (real-real similarity judgments) and their performance in test phase 1 (nonce-real similarity judgments).

2436 observations (across participants) were excluded for a log RT falling outside the Tukey interval for the participant that gave that response. The final dataset size after this filtering was 255964 observations (127430 nonce-real judgments) with each list having at least 18 response sets (out of the 30 total collected).

5.5 Results

As our measure of semantic information, we compute the Spearman correlation between the similarity judgments that participants gave for the novel word in test phase 1 and the similarity judgments for the true word that occurred in the utterances participants were trained on—i.e. the similarity judgments collected by White et al. (under review). Figure 8 shows the mean of these correlations by-participant correlations plotted by condition, along with 95% confidence intervals.

We see three trends in these data, which we first describe qualitatively, and then give statistical support for. The reason for giving this qualitative analysis first is that it is useful to get an overall picture of the variables of interest before adding in variables that control for the informativity of particular items and particular participants' semantic representations, which makes interpretation more complex (but still unambiguous).

5.5.1 Qualitative analysis. First, for all LEXICAL CONTEXT × CONTEXT CLASS pairs except *play real*, we find that getting more high informativity contexts improves learning relative to having fewer high information contexts, while having more low informativity contexts results in either no change—e.g. in the play contexts—or possible worse learning—e.g. in the dinner context—relative to having fewer high information contexts. This finding makes sense

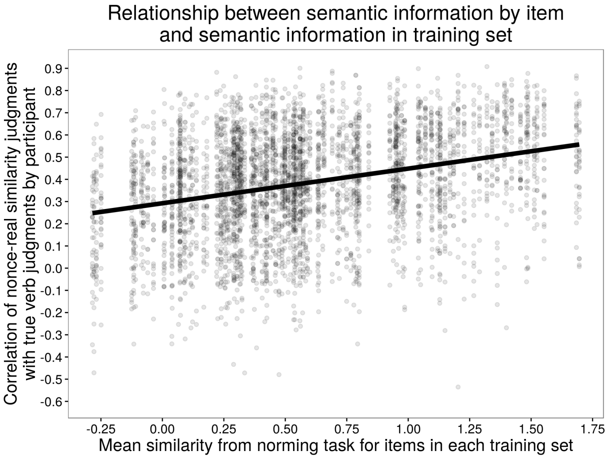

Figure 10. Relationship between mean similarity from norming task for items in each training set and the correlation between participants' performance in test phase 1.

in that getting more contexts that, in isolation, lead correct (or close-to-correct guesses) likely leads to more support for the correct guess; and getting more contexts that, in isolation, lead to incorrect guesses likely leads to simulated learnersgaining no new information or even compounding those incorrect guesses.

Second, the *play real* conditions show the opposite pattern. Indeed, getting more high informativity contexts with lexical information from the play context class actually hurts learning, while getting more low informativity contexts may improve learning (though only negligibly). One thing this may suggest is that the lexical information in play contexts is only really helpful when one has little data about a verb's distributional properties. And once one gets more distributional information about a verb, it is more helpful to rely on syntactic information. This is consistent with the pattern in the *play nonce* conditions, where more information is helpful.

Finally, we see that lexical information produces a substantial improvement in learning for the *dinner* contexts, both in the high and low informativity cases. This may suggest that semantic information in dinner contexts comes from both syntactic and lexical sources in combination, where in the play contexts, lexical information seems to be useful with little information, with syntactic information taking over as the amount of information grows.

One question that arises here is to what extent these effects are explainable by appealing to the informativity of particular items and to what extent they are about combining information from those items. This is what our quantitative analysis controls for.

5.5.2 Quantitative analysis. For our quantitative analysis, we begin with a linear mixed effects model with fixed effects for CONTEXT CLASS, LEXICAL CONTEXT, INFORMATIVITY, and TRAINING SIZE as well as all possible two-, three-, and four-way interactions among these variables, along with random intercepts for unique participant and true verb and by-verb random slopes for each of the terms in the fixed effects structure. We found convergence issues due to singular Hessians with the model containing the full random slopes that caused us to fall back to models with only the random intercepts. We refer to the resulting model as the BASE MODEL.

Prior to testing each of the fixed effects in the BASE MODEL, we test four potential control variables for inclusion. As the first potential control variable, we calculate the correlation between the similarity judgments for real verb-real verb pairs that participants gave in test phase 2 and the similarity judgments collected by White et al.. This correlation tells us how close to the population average each participant's similarity judgments for real verbs are. Figure 9 shows the test phase 1 correlations (y-axis), which we henceforth refer to as NONCE SIMILARITY, plotted against the test phase 2 correlations (x-axis), which we henceforth refer to as REAL SIMILARITY. The quantities are correlated at 0.295 (95% CI=[0.27, 0.32]), and a log-likelihood ratio test comparing the base model with one that also includes REAL SIMILARITY ($\chi^2(1) = 85.56, p < 0.001$) suggests that REAL SIMILARITY needs to be controlled for. We refer to the base model with REAL SIMILARITY added as the BASE+REAL SIMILARITY model.

As the second potential control variable, we calculate the mean accuracy in the norming tasks across items in the training set that each participant saw—i.e. the accuracies plotted in Figure 4. Adding this variable (TRUE ACCURACY) to the BASE+REAL SIMILARITY model did not result in a significantly better model by a log-likelihood ratio test ($\chi^2(1) = 0.82, p = 0.37$), and so we do not include it.

As the third potential control variable, we calculate the mean similarity to the true verb in the norming tasks across items in the training set that each participant saw—i.e. the similarities plotted in Figure 7. Adding this variable (TRUE SIMILARITY) to the BASE+REAL SIMILARITY model did result in a significantly better model by a log-likelihood ratio test ($\chi^2(1) = 40.81, p < 0.001$), and so we do include it. This yields the BASE+REAL SIMILARITY+TRUE SIMILARITY. Figure 10 plots the relationship between TRUE SIMILARITY and NONCE SIMILARITY.

Since TRUE SIMILARITY only captures similarity for inaccurate responses, we also consider the possibility that, even if the simple effect of accuracy is not significant, it may interact with the TRUE SIMILARITY variable. We thus compare the BASE+REAL SIMILARITY+TRUE SIMILARITY to one that also includes both the simple effect of TRUE ACCURACY and the TRUE SIMILARITY×TRUE ACCURACY interaction using a log-likelihood ratio test, finding that it is significant ($\chi^2(2) = 20.78, p < 0.001$). Thus, we retain this BASE+REAL SIMILARITY+TRUE SIMILARITY×TRUE AC-

CURACY model.

As the final potential control variable, we calculate the mean proportion of nonverb responses in the norming tasks across items in the training set that each participant saw—i.e. the nonverb proportions plotted in Figure 5. Adding this variable (NONVERB PROPORTION) to the BASE+REAL SIMILARITY+TRUE SIMILARITY×TRUE ACCURACY model did not result in a significantly better model by a log-likelihood ratio test ($\chi^2(1) = 2.23, p = 0.14$), and so we do not include it.

We now turn to the variables of interest. We first ask whether the four-way interaction among CONTEXT CLASS, LEXICAL CONTEXT, INFORMATIVITY, and TRAINING SIZE was significant by comparing the BASE+REAL SIMILARITY+TRUE SIMILARITY×TRUE ACCURACY model against the same model without this four-way interaction. We find that the four-way interaction is not significant by a likelihood ratio test ($\chi^2(1) = 1.47, p = 0.23$), and so we do not retain it.

Next, we test each of the four possible three-way interactions in the same way. We find that the three-way interactions among CONTEXT CLASS, LEXICAL CONTEXT, and INFORMATIVITY ($\chi^2(1) = 7.92, p < 0.01$) as well as CONTEXT CLASS, INFORMATIVITY, and TRAINING SIZE ($\chi^2(1) = 13.44, p < 0.001$) are significant by a log-likelihood ratio test, but the three-way interactions among CONTEXT CLASS, LEXICAL CONTEXT, and TRAINING SIZE ($\chi^2(1) = 2.21, p = 0.14$) and LEXICAL CONTEXT, INFORMATIVITY, and TRAINING SIZE ($\chi^2(1) = 1.03, p = 0.31$) are not. We test the model without both the latter three-way interactions against the one containing all of them, and find that both three-way interactions together are not significant ($\chi^2(2) = 3.20, p = 0.20$); thus, we retain only the former two three-way interactions.

This leaves only a single two-way interaction that is not a component of one of the retained three-way interactions: the two-way interaction between LEXICAL CONTEXT and TRAINING SIZE. Using the same testing procedure, we find that this two-way interaction is not significant ($\chi^2(1) = 0.43, p = 0.51$), and so we do not retain it. There are no further simple effects or interactions not contained within the retained three-way interactions, and so we analyze the coefficients of the resulting model.

Table 1 shows the coefficients for this model, which was fit using a dummy coding with *play* as the reference level for CONTEXT CLASS, *nonce* as the reference level for LEXICAL CONTEXT, *low* as the reference level for INFORMATIVITY, and *small* as the reference level for TRAINING SIZE. All continuous variables (REAL SIMILARITY, TRUE SIMILARITY, and TRUE ACCURACY) were standardized prior to fitting the model. This table contains the coefficient estimates for both the experimental variables and the control variables, which are greyed out. We walk through the significant experimental variables in turn.

But before moving forward, it is important to be clear on what these coefficients represent. The coefficient for each variable of interest is whatever effect a particular aspect of the experimental design had that cannot be explained purely by the informativity of particular items. That is, it specifies how much semantic information is added (or lost) by seeing those items together.

First, we see that the intercept, which corresponds to the left-most point in Figure 8, is significantly positive. This is unsurprising in light of Figure 8 and suggests that simulated learners that received even the small low informativity nonce training set from the play context show some ability to recover the true word's meaning.

The second significant effect we see is TRAINING SIZE. This suggests that increasing the training size in the small low informativity nonce training set increases participants' ability to recover the true verb's meaning over and above what we would expect from the informativity of those items. Note first that we see this increase even in light of the fact that the slope from the first to the third point in Figure 8 is negative. This suggests that participants are able to combine even the low play informativity contexts to better effect than might be expected.

The third significant effect is the interaction between LEXICAL CONTEXT and INFORMATIVITY, which is negative. This suggests that participants are not able to use the high informativity lexical information available in the play contexts to the extent that we would expect from (i) the informativity of those instances and (ii) the extent to which they were able to use the low informativity nonce cases. Consonant with what we noted in our qualitative analysis above, this may furthermore suggest that much of the useful information in the play contexts is actually syntactic in nature.

The fourth significant effect is the interaction between INFORMATIVITY and TRAINING SIZE, which is also significantly negative. This means that participants did worse in the large high informativity play condition than we might expect from (i) the item-level informativity alone and (ii) the extent to which they were able to use the low informativity nonce cases. This is also consonant with what we noted in our qualitative analysis above in that this may furthermore suggest that much of the useful information in the play contexts is available from very few learning instances.

The fifth and sixth significant effects are negative interactions between CONTEXT CLASS and INFORMATIVITY and between CONTEXT CLASS and TRAINING SIZE. This means that participants did worse in the nonce dinner conditions than we expect from item-level informativity in those conditions and their ability to use nonce items in the play conditions.

This contrasts with the seventh and eighth significant effects, which are positive three-way interactions between LEXICAL CONTEXT, CONTEXT CLASS, and INFORMATIV-

Table 1

Coefficients for final mixed effects model of NONCE SIMILARITY. *The model was fit using a dummy coding with* play *as the reference level for* CONTEXT CLASS, nonce *as the reference level for* LEXICAL CONTEXT, low *as the reference level for* INFORMATIVITY, *and* small *as the reference level for* TRAINING SIZE. *Control variables are greyed out.*

Term	**Estimate**	**Std. Err.**	
Fixed effects			
Intercept	0.39	(0.02)	***
LEXICAL CONTEXT	−0.01	(0.01)	
CONTEXT CLASS	0.00	(0.02)	
INFORMATIVITY	0.03	(0.02)	
TRAINING SIZE	0.04	(0.01)	***
REAL SIMILARITY	0.04	(0.01)	***
TRUE SIMILARITY	0.05	(0.01)	***
TRUE ACCURACY	0.01	(0.01)	
TRUE SIMILARITY × TRUE ACCURACY	−0.02	(0.00)	***
LEXICAL CONTEXT × CONTEXT CLASS	0.02	(0.02)	
LEXICAL CONTEXT × INFORMATIVITY	−0.03	(0.02)	*
INFORMATIVITY × TRAINING SIZE	−0.05	(0.02)	*
CONTEXT CLASS × INFORMATIVITY	−0.07	(0.02)	***
CONTEXT CLASS × TRAINING SIZE	−0.06	(0.02)	***
LEXICAL CONTEXT × CONTEXT CLASS × INFORMATIVITY	0.07	(0.02)	**
CONTEXT CLASS × INFORMATIVITY × TRAINING SIZE	0.09	(0.02)	***
Random effects	*Std. Dev.*		
PARTICIPANT	0.09		
VERB	0.06		

$p < 0.001$, $p < 0.01$, $^*p < 0.05$

ITY and between CONTEXT CLASS, INFORMATIVITY, and TRAINING SIZE. These three-way interactions effectively reverse the two-way interactions just mentioned. They suggest that, once participants get access to both lexical information and more data from the dinner context, they are better able to infer a verb's meaning. This can be seen in Figure 8 by how large the jump from nonce dinner contexts to real dinner contexts is. Thus, in contrast to the play contexts, where syntactic information seems to be the most useful, in dinner contexts, both syntactic and lexical information seems to be required.

5.6 Discussion

We have established two main results. First, context classes with less distributional complexity (e.g. play contexts) yield better learning when simulated learners have fewer samples from that distribution, and context classes with greater distributional complexity (e.g. meal contexts) yield better learning when those learners have more samples from that distribution. Second, participants (simulated learners) take advantage of only syntactic information in context classes with less distributional complexity, but they are only able to take advantage of the information in the higher complexity contexts if they have access to both syntactic and lexical information.

We take these results to be suggestive that context class plays an important role in determining how learners should update their hypotheses about a word's meaning given linguistic cues extracted from a particular context class. In particular, semantic information extracted from some context classes may be used to form initial imprecise hypotheses about a verb's semantics. This would make play contexts optimal for such coarse-grained learning of verbs' meanings, since extracting a good portion of the semantic information from linguistic cues found in play contexts does not require learners to know many lexical items. In contrast, semantic information extracted from other context classes may be used to refine those initial hypotheses. This would make dinner contexts optimal for such fine-grained learning, which may only happen after more lexical items have been learned.

6 General discussion

It is often tempting to identify distributional complexity with semantic information—probably because some of our best theories of information are fundamentally distributional in nature (Shannon, 1948). It is important, however, to remember that tracking distributional properties is merely a means to an end. We use distributional properties to learn; we don't use learning to obtain distributional properties.

Studying the mismatch between distributional complexity and semantic information makes this truism all the more apparent. The full range of context classes that children find themselves in no doubt have even more intricate and finely varying distributional structures than we have explored here. A major question for future research is how this variation undergirds children's language learning.

If our suggestion is correct that semantic information extracted from some context classes may be used to form initial imprecise hypotheses about a verb's semantics while semantic information extracted from other context classes may be used to refine those initial hypotheses, a potentially fruitful direction for future research involves understanding how learners know to treat semantic information extracted from a particular context class in a particular way. One possibility is that, rather than being an important source of semantic information in and of itself, the distributional complexity associated with a context class may act to trigger some control mechanism that determines how particular linguistic cues are used during learning. For instance, perhaps learners determine how much they should alter their hypothesis about a word's meaning based on the context class from which some linguistic cues were extracted.

We noted that one reason that meal contexts may contain higher distributional complexity is that these contexts mix child-directed speech with adult-directed speech in different proportions to toy play and book-reading contexts, which presumably show much higher levels of child-directed speech. But do these mixtures themselves show important sources of variations in children's input? And how might this variation interact with other factors, such as socioeconomic status, which is already known to have an effect on the distribution of particular kinds of structural complexity (cf. Hoff-Ginsberg 1991; Salo et al. 2015; Weizman and Snow 2001; Yont et al. 2003 among others)?

An important cross-cutting question is when and how children can use particular distributional properties of the input. There is growing evidence that children's input is not identical to their intake—i.e. the aspects of their input that they actually perceive and use in the course of learning (Lidz and Gagliardi, 2015; Omaki and Lidz, 2015). Intake can be modulated by a whole host of cognitive factors, such as memory constraints, predictive mechanisms, and aspects of the grammar that have already been acquired (Gagliardi, Feldman, and Lidz, 2016; Lidz, White, and Baier, under review; Medina et al., 2011; Omaki, 2010; Trueswell et al., 2013). And so another important question for future research is how these factors interact with the kind of modulation of distributional complexity and semantic information in the input we have found here.

Acknowledgments

We are grateful to audiences at the University of Maryland, Johns Hopkins University, IASCL 2014, and CUNY 2016 as well as Naomi Feldman, Hal Daumé III, Ellen Lau, Barbara Landau, John Trueswell, Roger Levy, Bob Frank, and Victor Ferreira for useful comments on this work. This work was supported by NSF BCS DDRIG grant 1456013 and NSF BCS grant 1124338, and White was supported by NSF DGE IGERT grant 0801465.

References

Arunachalam, Sudha, and Sandra R. Waxman. 2010. Meaning from syntax: Evidence from 2-year-olds. *Cognition* 114:442–446.

Baldwin, Dare A. 1991. Infants' contribution to the achievement of joint reference. *Child Development* 62:874–890.

Baldwin, Dare A. 1993. Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology* 29:832–843.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation* 43:209–226.

Bloom, Paul. 2000. *How Children Learn the Meaning of Words*. MIT Press.

Blum-Kulka, Shoshana. 1997. *Dinner Talk: Cultural patterns of sociability and socialization in family discourse*. New York: Lawrence Erlbaum Associates.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45:5–32.

Brown, Roger. 1957. Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology* 55:1.

Brown, Roger. 1973. *A First Language: The early stages*. Harvard University Press.

Bruner, Jerome. 1983. *Child's Talk: Learning to use language*. Oxford: Oxford University Press.

Bunger, Ann, and Jeffrey Lidz. 2004. Syntactic bootstrapping and the internal structure of causative events. In *Proceedings of the 28th Annual Boston University Conference on Language Development*, ed. Alejna Brugos, Linnea Micciulla, and Christine E. Smith, 74–85. Somerville, MA: Cascadilla Press.

Buttery, Paula. 2006. Computational Models for First Language Acquisition. Doctoral Dissertation, University of Cambridge.

Buttery, Paula, and Anna Korhonen. 2005. Large-scale analysis of verb subcategorization differences between child directed speech and adult speech. In *Verb Workshop 2005: Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, 1–6.

Ely, Richard, and Jean Berko Gleason. 1995. Socialization across contexts. In *The Handbook of Child Language*, ed. P. Fletcher and B. MacWhinney, 251–270. Oxford: Blackwell.

Ely, Richard, Jean Berko Gleason, Ann MacGibbon, and Elena Zaretsky. 2001. Attention to language: Lessons learned at the dinner table. *Social Development* 10:355–373.

Fisher, Cynthia, Henry Gleitman, and Lila R. Gleitman. 1991. On the semantic content of subcategorization frames. *Cognitive Psychology* 23:331–392.

Frank, Michael C., Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20:578–585.

Gagliardi, Annie, Naomi H. Feldman, and Jeffrey Lidz. 2016. Modeling statistical insensitivity: Sources of suboptimal behavior. *Cognitive Science* .

Gillette, Jane, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition* 73:135–176.

Gleason, Jean Berko. 1980. The acquisition of social speech routines and politeness formulas. In *Language: Social Psychological Perspectives*, ed. H. Giles, W.P. Robinson, and P.N. Smith, 21–27. Oxford: Pergamon Press.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition* 1:3–55.

Gleitman, Lila R., Kimberly Cassidy, Rebecca Nappa, Anna Papafragou, and John C. Trueswell. 2005. Hard words. *Language Learning and Development* 1:23–64.

Hoff-Ginsberg, Erika. 1991. Mother-child conversation in different social classes and communicative settings. *Child Development* 62:782–796.

Jones, Celeste Pappas, and Lauren B. Adamson. 1987. Language use in mother-child and mother-child-sibling interactions. *Child Development* 356–366.

Landau, Barbara, and Lila R. Gleitman. 1985. *Language and Experience: Evidence from the Blind Child*, volume 8. Harvard University Press.

Lederer, Anne, Henry Gleitman, and Lila Gleitman. 1995. Verbs of a feather flock together: semantic information in the structure of maternal speech. In *Beyond Names for Things: Young Children's Acquisition of Verbs*, ed. M. Tomasello and W.E. Merriman, 277–297. Hillsdale, NJ: Lawrence Erlbaum.

Leech, Kathryn A., and Meredith L. Rowe. 2014. A comparison of preschool children's discussions with parents during picture book and chapter book reading. *First Language* 34:205–226.

Lidz, Jeffrey, and Annie Gagliardi. 2015. How nature meets nurture: Universal grammar and statistical learning. *Annual Review of Linguistics* 1:333–353.

Lidz, Jeffrey, Aaron Steven White, and Rebecca Baier. under review. The role of incremental parsing in syntactically conditioned word learning. University of Maryland.

MacWhinney, Brian. 2014a. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

MacWhinney, Brian. 2014b. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Masur, Elise F., and Jean B. Gleason. 1980. Parent-child interaction and the acquisition of lexical information during play. *Developmental Psychology* 16:404.

Medina, Tamara Nicol, Jesse Snedeker, John C. Trueswell, and Lila R. Gleitman. 2011. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences* 108:9014–9019.

Naigles, L., Henry Gleitman, and Lila Gleitman. 1993. Syntactic bootstrapping and verb acquisition. In *Language and Cognition: A Developmental Perspective.*. Norwood, NJ: Ablex.

Naigles, Letitia. 1990. Children use syntax to learn verb meanings. *Journal of Child Language* 17:357–374.

Naigles, Letitia R. 1996. The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition* 58:221–251.

Nappa, Rebecca, Allison Wessel, Katherine L. McEldoon, Lila R. Gleitman, and John C. Trueswell. 2009. Use of speaker's gaze and syntax in verb learning. *Language Learning and Development* 5:203–234.

Omaki, Akira. 2010. Commitment and Flexibility in the Developing Parser. Doctoral Dissertation, University of Maryland.

Omaki, Akira, and Jeffrey Lidz. 2015. Linking parser development to acquisition of syntactic knowledge. *Language Acquisition* 22:158–192.

Papafragou, Anna, Kimberly Cassidy, and Lila Gleitman. 2007. When we think about thinking: The acquisition of belief verbs. *Cognition* 105:125–165.

Parisse, Christophe. 2000. Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, & Computers* 32:468–481.

Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61:127–159.

Roland, Douglas, and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Lin-*

guistics, volume 2, 1122–1128.

Roland, Douglas, and Daniel Jurafsky. 2002. Verb sense and verb subcategorization probabilities. In *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, ed. Paola Merlo and Suzanne Stevenson, 325–345. John Benjamins Publishing Company.

Rowe, Meredith L. 2013. Decontextualized language input and preschoolers' vocabulary development. In *Seminars in Speech and Language*, volume 34, 260–266. Thieme Medical Publishers.

Sagae, Kenji, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2007. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, 25–32. Association for Computational Linguistics.

Salo, Virginia C., Meredith L. Rowe, Kathryn A. Leech, and Natasha J. Cabrera. 2015. Low-income fathersâĂŹ speech to toddlers during book reading versus toy play. *Journal of Child Language* 1–15.

Shannon, Claude E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27:379–423, 623–656.

Snedeker, Jesse. 2000. Cross-situational observation and the semantic bootstrapping hypothesis. In *Proceedings of the thirtieth annual child language research forum. Stanford, CA: Center for the Study of Language and Information*.

Snedeker, Jesse, and Lila Gleitman. 2004. Why it is hard to label our concepts. In *Weaving a Lexicon*, 257–294. Cambridge, MA: MIT Press.

Snedeker, Jesse, Lila Gleitman, and Michael Brent. 1999. The successes and failures of word-to-world mapping. In *Proceedings of the Twenty-third Boston University Conference on Language Development*, ed. A. Greenhill, M. Hughs, and H. Walsh. Citeseer.

Tomasello, Michael. 1992. The social bases of language acquisition. *Social Development* 1:67–87.

Trueswell, John C., Tamara Nicol Medina, Alon Hafri, and Lila R. Gleitman. 2013. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology* 66:126–156.

Weizman, Z. O., and C. E. Snow. 2001. Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology* 37:265–279.

White, Aaron Steven, Valentine Hacquard, and Jeffrey Lidz. under review. Semantic information and the syntax of propositional attitude verbs. University of Maryland.

Yont, Kristine M., Catherine E. Snow, and Lynne Vernon-Feagans. 2003. The role of context in mother-child interactions: An analysis of communicative intents expressed during toy play and book reading with 12-month-olds. *Journal of Pragmatics* 35:435–454.

Yu, Chen, and Linda B. Smith. 2007. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science* 18:414–420.

Yu, Chen, and Linda B. Smith. 2012. Modeling cross-situational word-referent learning: Prior questions. *Psychological review* 119:21.

Yuan, Sylvia, and Cynthia Fisher. 2009. "Really? She Blicked the Baby?" Two-Year-Olds Learn Combinatorial Facts About Verbs by Listening. *Psychological Science* 20:619–626.

Yurovsky, Daniel, and Michael C. Frank. 2015. An integrative account of constraints on cross-situational learning. *Cognition* 145:53–62.