Shifting senses in lexical semantic development

Hugh Rabagliati[1], Gary F. Marcus[1], Liina Pylkkänen[1,2]

[1] Department of Psychology, New York University

[2] Department of Linguistics, New York University

Please address correspondence to:

Hugh Rabagliati

Department of Psychology

New York University

6 Washington Place, 8th Floor

New York, NY 10003

Tel: (212) 998-3870

Email: hugh@nyu.edu

Abstract

Most words are associated with multiple senses. A DVD can be round (when describing a disc), and a DVD can be an hour long (when describing a movie), and in each case DVD means something different. The possible senses of a word are often predictable, and also constrained, as words cannot take just any meaning: for example, although a movie can be an hour long, it cannot sensibly be described as round (unlike a DVD). Learning the scope and limits of word meaning is vital for the comprehension of natural language, but poses a potentially difficult learnability problem for children. By testing what senses children are willing to assign to a variety of words, we demonstrate that, in comprehension, the problem is solved using a productive learning strategy. Children are perfectly capable of assigning different senses to a word; indeed they are essentially adult-like at assigning licensed meanings. But difficulties arise in determining which senses are assignable: children systematically overestimate the possible senses of a word, allowing meanings that adults rule unlicensed (e.g., taking *round movie* to refer to a disc). By contrast, this strategy does not extend to production, in which children use licensed, but not unlicensed, senses. Children's productive comprehension strategy suggests an early emerging facility for using context in sense resolution (a difficult task for natural language processing algorithms), but leaves an intriguing question as to the mechanisms children use to learn a restricted, adult-like set of senses.

KEYWORDS: lexical development; lexical semantics; word learning; language acquisition

1.1 Introduction

Human language is filled with the ambiguous and non-literal. When the witches of Macbeth urge that their *fire burn and cauldron bubble*, they don't mean for the cauldron itself to melt and boil, but the magic potion inside. Similarly, if I *order some Beethoven from the music store*, I have not ordered a lump of the composer, but rather some of his works; if I *find the CD to be moving*, it is the composer's works that cause emotion, not the plastic CD itself. In each example, the surface meaning of the sentence seems implausible but by shifting the meaning of a critical constituent we can derive a reasonable interpretation.

However, the elasticity of meaning only stretches certain ways. Although *Beethoven* can refer to the composer's music, his music cannot refer to him; it is nonsensical to say that *the 8th symphony was deaf*. Similarly, *the CD* can refer to the composer's work, but not vice versa (e.g., *the 8th symphony was shiny*).

The child learning a language has to figure out these ground rules, a task that is far from trivial. Computer scientists have spent forty years failing to create a computer program that can adequately determine the similar but different meanings associated with words such as *Beethoven* or *CD*, which linguists call senses (for an overview see Miller, 1999). Yet by adulthood our ability to resolve a word's sense is extremely accurate. How, then, do children learn the ways a word's sense can change?

1.2 Shifts and senses

Although children eventually attain a remarkable degree of mastery over the scope and limits of word meaning, relatively little is known about how and when they

manage to do so. We know a great deal about how children use logical principles, theory of mind, syntax and other factors to determine the basic referent of a word when heard for the first time, but little about how a child's understanding of a word extends beyond those first encounters. Carey and Bartlett (1978) argued that any word requires a long period of slow discovery before a child finally determines its exact meaning, a point that has been echoed by Murphy (2001). Presumably, learning the many ways a word's meaning changes is similarly difficult.

What sorts of representations do children have to acquire? One possibility is that the lexicon simply lists a set of word forms paired with their meanings. Entries for words such as *DVD* would contain the form alongside a sense referring to a shiny disc and a sense referring to the movie stored on that disc. This theory is attractively simple and to some extent may even be true: there is no reason why frequently used senses of a word could not be stored together. But it has difficulty accounting for any sort of creative word use. For example, an animal's name can also refer to the food produced from it (compare *noisy chicken* and *tasty chicken*), and this sort of template generalizes across words with similar meanings. A reader offered a steaming saucer of *Sasquatch*[1] would find the food sense entirely transparent yet also entirely novel. If they had lacked productive means for changing a word's sense then the only available meaning would have been the often-encountered ANIMAL sense, and not the novel FOOD one (for discussion see Murphy, 2007).

[1] The sasquatch, also known as Bigfoot, is an alleged, ape-like creature inhabiting the Pacific Northwest of the USA.

Mastering this type of creativity requires the child to learn a number of productive shifts[2] that can generate senses based on certain aspects of a word's meaning. As shown below, we can use the name of an object to refer to its abstract contents (1), take a container's name to stand for its contents (2), or interpret a physical object as taking part in some unspecified event (3).

(1) The DVD was an hour long.

= *The movie on the DVD was an hour long.*

(2) The pot was stirred.

= *The content of the pot was stirred.*

(3) The boy began the book.

= *The boy began reading/writing the book.*

Under most theories, productive shifts in meaning are the result of rules applying over coherent semantic classes, such as containers or animals. For instance, a container-content rule takes a container meaning and transforms it to a contents meaning. These rules are not typically associated with any overt syntactic marker (contrast *The pot was washed* and *The pot was stirred*), so to ascertain whether a rule is being used requires the listener to judge which meaning is more plausible. The exact operations by which such rules occur is subject to debate; in particular it is not clear if shifts are lexical, syntactic, semantic or pragmatic phenomena (for discussion see Brennan & Pylkkänen, 2008;

[2] We use the term 'shift' to describe the process underlying a range of productive lexical phenomena, which go under labels as varied as polysemy, metonymy, coercion, systematic polysemy, deferred interpretation, sense transfer and more.

Copestake & Briscoe, 1995; Harris, Pylkkänen, McElree, & Frisson, 2008; Lapata & Lascarides, 2003; McElree, Traxler, Pickering, Seely, & Jackendoff, 2001; Miller, 1995; Murphy, 2007; Papafragou, 1996; Pustejovsky, 1995; Pylkkänen, 2008; Pylkkänen, Llinas, & Murphy, 2006; Pylkkänen & McElree, 2006, 2007). In addition, not every theory agrees that shifts require a system that is rule-based (Nunberg, 1979 ; 1995, 2004). But whatever the form of the theory, the child still needs to acquire a system that is productive.

While the child needs to learn which shifts to make, they also need to learn which shifts not to: Not every shift is possible. For example, although each shift in (1-3) above is licensed, shifting in the opposite direction is not possible. *DVD* has a sense similar to movie, but *movie* cannot be shifted to 'DVD the movie is on' to derive a plausible reading for (4). *The soup* is not easily shifted to its container in (5), and even though we interpret *Book* as 'Reading the book' in (3), we are unable to perform the same shift in reverse (6). Furthermore, there is limited cross-linguistic work demonstrating that the set of licensed shifts varies across languages. For instance, Kamei & Wakao (1992) argue that the producer-product shift (*Barty read Dickens*) is not licensed in Mandarin Chinese.

(4) The movie was shiny.

≠ *The DVD containing the movie was shiny.*

(5) The soup was cracked.

≠ *The pot containing the soup was cracked.*

(6) Reading the book was 200 pages.

≠ *The book that was read was 200 pages.*

1.3 Shift acquisition

How might children learn the set of licensed shifts while excluding the unlicensed ones? Because theories of lexical development (e.g., Bloom, 2000) assume words are form-meaning pairs, they cannot provide an adequate answer. If children were to learn word senses piecemeal, then each novel sense should be equally plausible, and this is clearly not the case: some novel senses (e.g., *saucer of Sasquatch*) are easily interpretable, but others (e.g., *shiny movie*) are not. Children, then, have to learn a productive system with a set of constraints on what makes a sense licensed. This means that the learner will face problems that are analogous to those encountered learning similarly generative systems, like syntax.

The foremost difficulty for any theory of syntactic development is the projection problem: avoiding the acquisition of an overly general grammar (Baker, 1979; Braine, 1971; Pinker, 1984, 1989). As an example, when learning about argument structure the child cannot simply assume that every verb undergoes passivisation. *Billy owns the books* alternates with *The books are owned by Billy*, but *Billy has the books* does not alternate with **The books are had by Billy*. At the same time, if children only passivise verbs that they have previously heard modeled as passives, they would never acquire an adult-like ability to generate new passives, such as *The message was tweeted by Stephen*. So, as with shifts, the learner has to acquire a system with a restricted level of productivity. Many theories of these restrictions assume that they are semantically defined: verbs with similar meanings have similar argument structure restrictions, and the nature of that

meaning determines the restrictions (e.g., Levin, 1993; Pinker, 1989; Pinker, Lebeaux, & Frost, 1987).

Much the same is true for shifts. Because only a subset of potential shifts is licensed, and semantic criteria appear to define those (e.g., words for containers can enter container-contents shifts), learning will be similarly difficult in lexical semantic development. Given this similarity, learning strategies that are successful in syntactic development may also aid lexical semantic development.

A crucial distinction for theories of the child's solution to the syntactic projection problem is whether they act as a conservative or productive learner (Baker, 1979; Bowerman, 1987; Braine, 1971; Pinker, 1979, 1984, 1989; Tomasello, 1992, 2003). A productive learner will infer a grammar that licenses the use of words in ways or constructions that have never been attested, and as a consequence this grammar may be incorrect. A conservative learner is unwilling to make inferences about the behavior of a constituent without direct positive evidence. They avoid the problems and pitfalls of over-generalization by not generalizing, a strategy that leaves them in danger of underestimating the grammar. A natural stepping-off point for assessing lexical semantic development is to ask whether children solve this projection problem as productive or conservative learners, that is to say, assessing whether they willingly make shifts that are not attested in the input.

In the remainder of this introduction, we set out evidence for whether children take a conservative or productive approach to learning about other semantic/pragmatic phenomena. Our experiments then test exactly which strategy children take for shifts. Productive children should willingly entertain a range of senses for each word, both

licensed and unlicensed. During comprehension they should use a range of contextual cues to search for the set of senses that, when assigned to the words of a sentence, best describe the current situation, whether these senses are licensed or unlicensed. We describe this strategy as one of 'situational fit'. In contrast, conservative children will have a bias against assigning additional senses to a word, which will be reflected in their comprehension/production as an unwillingness to use words in ways that have not previously been attested.

1.4 Conservativity and productivity in shift development

What reasons might children have to adopt a conservative or productive strategy? One major factor is the extent to which they wish to rely on vague context to determine the meanings of words. The evidence for a sense is typically ambiguous, because shifts lack valid cues such as morphosyntactic marking (in contrast with argument structure acquisition, where the presence of a syntactic structure like the passive is obvious from its constituents and their surface order). For example, *The DVD was good*, can be equally well interpreted as describing a disc or a film. Because context is typically ambiguous, the child will frequently be unable to say with certainty what meaning a word should take. A productive child would err on the side of context, and a conservative child should err on the side of prior experience.

The potential benefits of conservative and productive strategies can be seen in models of sense resolution from computational linguistics. Many modern Bayesian approaches aim to make as much use of context as possible (Manning & Schütze, 2002), per productive learners, but researchers have also had success with algorithms that

choose the dominant (most frequent) sense of a word and ignore anything other than highly unambiguous contexts (e.g., McCarthy, Koeling, Weeds, & Carroll, 2004), akin to conservative learners. In particular, conservative algorithms are particularly successful when information about contexts is highly uncertain; exactly the situation children find themselves in.

These two classes of algorithms are ultimately too extreme as theories of human lexical development: a child who completely ignored context would never be able to learn any shifts, and a child who relied on it would overestimate the set of shifts. However, the parallel should be clear: productive children should weight context heavily, while conservative children should weight it lightly. In the absence of any detailed theories on what restrictions on shifting exist, and therefore what constraints children might use, the productive-conservative distinction provides a first step to understanding this important aspect of lexical development.

1.4.1 Evidence for productivity

Across a variety of domains, the balance of evidence suggests that children adopt productive learning strategies during language acquisition. For argument structure, there is evidence for conservativity very early in development (around two years, Tomasello, 1992; Tomasello, 2003), but by approximately 4 years most children appear to be taking a productive approach, producing both novel passives (e.g.. How was it shoelaced?, Clark, 1982) and unlicensed passives (I don't like being falled down on!, Wasow, 1981). Pinker et al. (1987) argue that children between 4- and 7-years learn via a productive strategy that is mildly constrained: children will passivize essentially any verb, but are

more willing to passivize verbs that do not violate proposed restrictions on the passive (e.g., Jackendoff's Thematic Hierarchy Condition (Jackendoff, 1972), or Pinker's Thematic Core theory (Pinker, 1989)). Whether similar constraints exist for shifts is unclear.

Children of a similar age also appear to have little-to-no difficulty interpreting phenomena that are more clearly shift-like. Srinivasen and Snedeker (in prep) found that children can use both senses of abstract object containers, such as *DVD, while* Barner and Snedeker (2005, Experiment 3) show that 4-year-old children are able to use plural morphology (*some paper/some papers*) to determine the sense of a count/mass ambiguous noun. Both these results suggest that children can use familiar senses but do not necessarily speak to their learning strategy. Better evidence for productivity is provided by Bushnell & Maratsos (1984), who argued that children as young as two have little difficulty interpreting innovative class extensions, in which a word is used in a novel lexical category (e.g., *Can you lipstick the trashcan?*). Clark (1982) also documents a number of class-extensions in the speech of young children.

Children also appear adept at determining whether sentences have generic (*dogs have tails*) or non-generic readings (*All dogs have tails*). Gelman's extensive research on this suggests that a distinction between the two types of meaning is drawn early, before 4 years-of-age, and that children have little difficulty switching between forms in comprehension and production of both familiar and novel nouns, consistent with a productive learning strategy (Brandone & Gelman, 2009; Gelman & Bloom, 2007; Gelman, Goetz, Sarnecka, & Flukes, 2008; Gelman & Raman, 2007).

In addition, metaphor comprehension seems to be relatively productive (Vosniadou, 1987), contrary to claims made by Piaget (1926). The only limits on productive metaphor comprehension appear to result from children's knowledge of the domains described by the metaphor. Keil (1986) demonstrated that kindergartners were able to comprehend and explain novel metaphors like *The car was thirsty* where both the target domain (the car) and the source domain (the property of being thirsty) came from ontological categories that were familiar to the child (for example, animate and inanimate objects).

However it is not clear that the productivity seen in processing mass/count ambiguities, generic statements or metaphors should extend across all lexical semantic domains. Metaphors, although clearly related to shifts, may be an inaccurate comparison because shifts change the referent of a word, while metaphors describe a non-literal way of conceptualizing that word (see Lakoff & Johnson, 1980; Ortony, 1979).

Children's facility with mass/count ambiguity, class extensions and generics may also fail to generalize to shifts, because in each case syntactic cues (such as the presence or absence of determiners, quantifiers or plural morphemes) often overtly indicate exactly what type of meaning a noun phrase takes. It may be that children are more conservative when processing semantic phenomena that lack morphosyntactic markers, like shifts. Cimpian and Markman (2008) provide some confirmation of this, showing that 3-year-old children easily derive generic meanings indicated by syntactic cues but are less able to use a range of non-syntactic cues.

1.4.2 Evidence for conservativity

Consistent with an important role for syntactic cues, children appear to take a conservative approach to the acquisition of another lexical semantic phenomenon that lacks syntactic marking, scalar implicatures. In a scalar implicature, such as the inference from *John ate two candies* to *John did not eat three candies*, a listener decides that the set of meanings a speaker intends his utterance to describe is smaller than his utterance actually describes. For example, the sentence *John bought some candy at the store* is true if John bought a little bit of candy, but also (technically) if John bought all the candy in the store. Nevertheless adults strongly disprefer, or even rule unacceptable, the latter reading- they prefer *some* to mean *some but not all*. This preference is typically explained as the result of an implicature applied to the scalar term *some*: To say *some* when the speaker meant *all* would not be completely informative, so the speaker is assumed to have implied *some but not all*. Although the meanings derived from an implicature are not typically considered senses, there are clear parallels between the two. The two meanings considered in the computation of an implicature are clearly related, and the listener must use context to choose between them.

As demonstrated by Noveck (2001) and Papafragou and Musolino (2003), children as old as 10 years fail to compute implicatures when interpreting scalar terms, allowing both of the readings of *John bought some candy*, the scenario where John bought a tiny bit of candy and the pragmatically anomalous case in which John bought all the candy.

Although children are in some way overly liberal in allowing the pragmatically anomalous interpretation, these data have usually been argued to reflect a *conservative* development of lexical semantics. Noveck (2001) suggests that children understand that

the true meaning of *some* is *at least one*, and are unable or unwilling to apply the implicature operation in order to change this sense to *some but not all*. The implied conservativity means children are more logical than adults, in that they follow the truth conditions of a sentence, and fail to use any pragmatic filters upon it. To behave in this way, children would have to be extraordinarily conservative, because such meanings have almost certainly been frequently demonstrated in their input.

1.4.3 The current experiments

Here, we test if children are productive enough to change the truth conditions of a sentence using a shift. The truth conditions of a sentence containing a shift are often disjoint from its unshifted truth conditions, as in *the DVD was an hour long*, where it is the movie, not the disc, which is an hour long. If children are conservative in shifting between senses, they should find this sentence uninterpretable (because discs do not have temporal extent). If children are productive, they should understand that *DVD* requires its movie sense to assign a reasonable interpretation to the sentence. In the first experiment we test how willing children and adults are to make licensed and unlicensed shifts during comprehension, and how this changes over development, while Experiment 2 contains an additional test of shift production.

2 Experiment 1

Following work on semantic development by Keil (1979), adults and children aged 3.5 to 8.5 years were asked a series of predicability questions (*Could an X be Y?*) by an uninformed robot that they were trying to educate during a game. By manipulating both

the properties and whether the argument's sense could be altered via a shift, we derived a set of questions whose answers would vary depending on the participants' willingness to make licensed and unlicensed shifts.

Without a shift, match questions like (7) and (8) should be affirmed, and mismatch questions like 9 and 10 denied. The licensed mismatch question (9) could be affirmed using a licensed shift (for example, by changing the meaning from DISC to FILM). However, the unlicensed mismatch question (10) could only be affirmed using an unlicensed shift (in this case from FILM to DISC).

(7) Match: Could a DVD be round?

= *Could a DVD disc be round?*

(8) Match: Could a movie be an hour long?

= *Could a movie presentation be an hour long?*

(9) Licensed Mismatch: Could a DVD be an hour long?

= *Could a movie on a DVD be an hour long?*

(10) Unlicensed Mismatch: Could a movie be round?

≠ *Could a DVD containing a movie be round?*

We used predicability judgments rather than the standard truth value judgment task (Crain & Thornton, 2000; Gordon, 1996) because the context that is required to generate a shift often results in sentences with an interpretation that lacks a truth-value (e.g., in *The DVD was an hour long*, physical objects have no temporal length). If children cannot use shifts (e.g., they interpret *DVD* as a physical object), it is not clear how they could assign a truth value to such an utterance, and so questions about its truth or falsity would

be ill posed.

By contrast, a predicability question is always interpretable, even if the predication is nonsensical. Predicability questions and the truth value judgment task also have relatively similar task demands, as in each case children need to make judgments of semantic plausibility, rather than judgments of structural acceptability. The one major concern over predicability questions is whether they are appropriate for very young children, either because they do not understand the task, or because they suffer from a response bias. Keil (1979) was able to record predicability judgment from children as young as 3, suggesting the task can feasibly be applied. In addition, we included warm-up trials in which the experimenter emphasized how the questions could be silly, as well as several analyses to control for response bias.

To test which learning strategy children used, we examined the acquisition profiles of three "shift templates:" shifts between items that fall into different ontological categories. In particular, we used licensed/unlicensed container-contents shifts which shifted between objects and substances (e.g., *Could a pot be stirred?/ *Could some soup be cracked?),* object-event shifts (e.g., *Could a boy begin a book?/ *Could reading a book be little?*) as well as object-abstract shifts (e.g., *Could a DVD be an hour long?/ *Could a movie be round?).* Children applying a conservative learning strategy will initially answer "No" to any predicability questions that require an unlicensed shift to be resolved as well as questions requiring any unfamiliar but licensed shifts, but over development will affirm more questions requiring licensed shifts. By contrast, productive learners will initially affirm both licensed and unlicensed shifts, and over development will increasingly reject questions requiring unlicensed shifts.

2.1 Method

2.1.1 Participants

53 children aged between 3.5 and 8.5 years participated. Mean age was 6;7, (SD = 18 months). 18 college-age adults also participated. All subjects spoke English as their first language. Rather than rely on chronological age as a predictor of ability, we determined each child's linguistic age using the TELD 3 standardized test (Hresko, Reid, & Hammill, 1999). Children had a mean linguistic age of 6;6 (SD = 16 months).

Rather than binning subjects into age groups, we analyzed all data using regressions that treated linguistic age as a continuous variable. Treating age continuously is both theoretically and practically advantageous. It is theoretically advantageous because age is clearly not discrete, and binning subjects by year imposes arbitrary distinctions (e.g., a subject aged 6;11 will be grouped with a subject aged 6;0, but apart from a subject aged 7;0). It is practically advantageous because binning subjects into age blocks (or binning any continuous variable) results in a considerable loss of power (MacCallum, Zhang, Preacher, & Rucker, 2002; Selvin, 2004; van Walraven & Hart, 2008) due to the loss of information in the predictor. In light of this, our number of subjects per year-of-life was lower than in some other developmental studies (although it was similar to previous studies on similar phenomena, e.g., Pinker et al, 1987).

One potential problem with treating age as continuous in a linear regression is the possibility of missing unexpected developmental patterns, such as U-shaped development, which might be uncovered using discrete predictors. Our hypotheses about learning strategy all predict changes in answers over age to be monotonic increasing

(conservative) or decreasing (productive). As a safe guard against the possibility of a more complicated pattern, we also graph our data split into four age groups, below 5 years (range: 3;6 to 5;0 years, n = 19), 5;0- to 7;0-years (n = 16), above 7-years (range: 7;1 – 8;6, n = 19), and adults (n = 18), which should reveal any unexpected developmental patterns.

2.1.2 Materials

To create sets of predicability questions we used pairs of words that fit into the roles of each licensed/unlicensed shift template, as in (7) to (10). For example, a *DVD* is an object and, because of the object-abstract shift, it can be shifted to a FILM sense in a licensed mismatch questions (9). By contrast a *movie* is an abstract object and, because the reverse shift is unlicensed, it cannot be shifted to a DISC sense in an unlicensed mismatch question (10).

We used 5 quadruplets of questions per template (see Table 1). For the object-event shift, an additional fifth cell containing a fully specified *begin*-phrase was added (e.g., *Could a boy start drawing a picture?*). Since this condition produced the same results as the other licensed mismatch question, this data is not used further in the current paper[3].

We were careful to control for contamination by a Yes-bias. To gauge the baseline probability of affirming or rejecting a question due to noise or response bias, rather than shift use, we created an additional set of 8 control questions (4 match and 4 mismatch) that we judged to be extremely difficult to shift. We used animacy violations (such as *Could a rock be angry?*), as these did not fit any known shift template, and are drawn

[3] Experiment 2 excluded this condition, and found an essentially identical result, so its inclusion here did not seem to influence participants' answers.

from a domain (the animate-inanimate distinction) which children are highly knowledgeable about (Keil, 1979).

In addition, we wanted to ensure that children did not only affirm or deny questions because of the frequency with which their predicates and arguments co-occur. We used latent semantic analysis to control for co-occurrence: LSA cosines were either equal between licensed/unlicensed pairs, or the cosine (and therefore co-occurrence) was greater for the unlicensed question.

Finally, a picture was created for each set of questions, depicting its arguments, which was presented to the children by the robot alongside its questions, to make processing of the query easier. For example, the picture for the DVD-movie set depicted the disc, and its case displaying a scene from a potential movie (a picture of the jungle).

Table 1: Example stimulus items for each question type and shift template. LSA cosine indicates the text co-occurrence between predicate and argument.

	Licensed Match Question	**Licensed Mismatch Question**	**Unlicensed Match Question**	**Unlicensed Mismatch Question**
Container-Contents	Could a pot be cracked?	Could a pot be stirred?	Could some soup be stirred?	Could some soup be cracked?
LSA Cosine	*0.15 (0.08)*	*0.24 (0.1)*	*0.15 (0.09)*	*0.28 (0.05)*
Object-Event	Could a picture be large?	Could a boy start a picture?	Could drawing a picture be quick?	Could drawing a picture be large?
LSA Cosine	*0.15 (0.06)*	*0.14 (0.06)*	*0.16 (0.05)*	*0.14 (0.07)*
Object-Abstract	Could a DVD be round?	Could a DVD be an hour long?	Could a movie be an hour long?	Could a movie be round?
LSA Cosine	*0.08 (0.09)*	*0.07 (0.03)*	*0.11 (0.02)*	*0.21 (0.07)*

The robot's questions were recorded in a neutral tone by a female North-American English speaker, and presented over 5 blocks. Each block contained a full set of items from each shift template with presentation order determined by pseudo-randomization, under the constraints that order of mention of licensed and unlicensed shifts was counter-balanced over blocks, and a question could not immediately follow another question from the same shift template. Each block was separated by a control question. All materials are given in Appendix A. Because the current experiment was included as part of a larger battery of tests for individual differences in language and cognitive development, only one presentation order was used, consistent with standard individual differences testing procedures (see Carlson & Moses, 2001). We believe the current results are interesting enough to present in isolation from the other tests, but the procedure does introduce the unfortunate possibility of an order effect. However, because the major results of the experiment are replicated in Experiment 2, where order of presentation was varied, it is unlikely that order provided an important contribution to our results.

2.1.3 Procedure

Participants made their predicability judgments under the guise of a game, answering an uninformed robot's questions about the world (in a procedure partly drawn from Fernandes, Marcus, Di Nubila, & Vouloumanos, 2006). Adults were told that the game was designed for children, but otherwise the procedure was essentially identical. At the start of the game, an experimenter explained that the robot would be asking the child about the world, and that, because the robot did not know very much, it would sometimes

ask very silly questions. The child was then given two warm-up trials with the experimenter, who again emphasized possible silliness of questions.

On each trial, the robot (controlled by a confederate) asked a predicability question while displaying the picture associated with the question on two computer monitors, disguised as her eyes. Participants responded with a yes or no answer, recorded by the experimenter. If the participant requested, the robot would repeat each question once, as would the experimenter if necessary. The experimenter also explained any terms that participants did not understand. The experimenter often stressed that participants should be sure to tell the robot whenever she said anything silly, but did not give differential feedback based on participants' answers. All sessions were videotaped.

2.2 Analyses and Results

We assessed learning strategy through a two-stage analysis. First, we assessed changes in the acceptance rates of licensed and unlicensed mismatch questions. Second, we investigated the possibility of response bias effects through both a comparison with animacy violation control questions, and a *d'* analysis; the latter also allows us to control for vocabulary differences between the stimuli.

All analyses were performed using regressions, allowing us to treat age as a continuous variable. To provide a clear sense of how children's data varied over age, each analysis includes a graph of the regression estimates, alongside a graph that breaks the data down by both the shift template and the four age groups discussed in section 2.1.1. Finally, to provide some indication of the extent of individual variation in the data, we have also provided histograms of individual *d'* scores for the different conditions in

the supplementary materials.

2.2.1 Acceptance rates for licensed and unlicensed mismatch questions

Under a conservative learning strategy, young children should be unwilling to accept any unlicensed mismatch questions, and the probability of accepting a licensed mismatch question should increase with age. With a productive learning strategy, young children should initially accept both licensed and unlicensed mismatch questions, and reject unlicensed questions as they age.

The bar charts in Figure 1 display the proportion of Yes answers for each condition[4] (i.e., affirming that X could be Y). Visual inspection suggests that children employ a productive learning strategy. The proportion of Yes answers to licensed mismatch questions was high, and essentially constant across the four age groups, as would be predicted if all groups were making shifts while interpreting the question. Children's responses to *unlicensed* mismatch questions, however, differed from adults, with important differences between the different templates. In the container-contents template, children's answers to unlicensed mismatch questions were again essentially adult-like, with a low proportion of Yes answers that did not greatly differ between the age groups. But for the object-event and object-abstract templates, children, especially younger children, appeared distinctly non-adult-like. In particular, they more often accepted unlicensed mismatch questions than adults, again consistent with a productive learning strategy

To analyze these apparent effects we employed mixed-effects logistic regressions,

[4] As well as Control questions from Section 2.2.2.

which are more appropriate for analyzing categorical data than ANOVAs[5]. For each shift template, answers to mismatch questions were entered into a model with fixed effects of language age[6], mismatch question type (dummy coded with 0 = licensed, 1 = unlicensed), and an interaction between the two, as well as a random intercept and effect of question type for subjects, and a random intercept and effect of language age for items. We fit all models using the lmer function of the lme4 package (Bates & Sarkar, 2008) in R.

Based on this regression, Figure 2 displays an estimate of the proportion of Yes answers for each condition over age, where the solid line represents the estimated proportion of Yes answers to licensed mismatch questions over age, and the dotted line is the estimated proportion of Yes answers to unlicensed mismatch questions. In addition, the points in Figure 2 indicate the raw Yes and No answers for each condition, jittered around 1 and 0 respectively. The relevant terms of the regressions are described below, while the full regressions are included in Table 1 of the supplementary materials. N for each regression was 710, while deviances were: container-contents = 695, object-event = 508, and object-abstract = 697.

Our regressions confirmed the qualitative description of the effects given at the start of this section. Children behaved in an adult-like manner when answering licensed questions: The proportion of *Yes* answers provided did not change over age (Object-Event: β = -0.001, s.e. = 0.003, Wald Z = 0.24*, ns;* Object-Abstract: β = 0.01, s.e. = 0.006, Z = 1.38, *ns,* Container-Contents: β = -0.001, s.e. = 0.002, Z = 0.25 *ns*).

[5] For a discussion of the problems with ANOVA and categorical data, and an introduction to logistic regressions, see Jaeger (2008).

[6] Language age was standardized by subtracting the age of the youngest child in our dataset from all other ages in order that the intercept was at 44 months, causing other predictors to estimate their effects at the youngest age, and not extrapolated to 0 months.

For the container-contents template alone, children also behaved in an adult-like manner when answering unlicensed questions: They were less likely to affirm these unlicensed questions than their paired licensed questions (β = -2.33, s.e. = 0.51, Z = 4.59, $p < .01$), and the affirmation rate did not change over age, indicated by a non-significant interaction between age and question type (β = -0.009, s.e. = 0.005, Z = 1.90, p = .06).

Finally, for the object-event and object-abstract templates, our regressions confirmed that younger children were distinctly non-adult-like in their answers to the unlicensed questions. In particular, a reliable interaction between age and question type indicated that young children accepted a higher proportion of unlicensed mismatch questions than adults, consistent with a productive learning strategy. (Object-Event: β = -0.02, s.e. = 0.006, Z = 3.75, $p < .01$; Object-Abstract: β = -0.022, s.e. = 0.008, Z = 2.95, $p < .01$). This age-related decrease in acceptance rate for unlicensed questions can clearly be seen in the regression estimates of Figure 2.

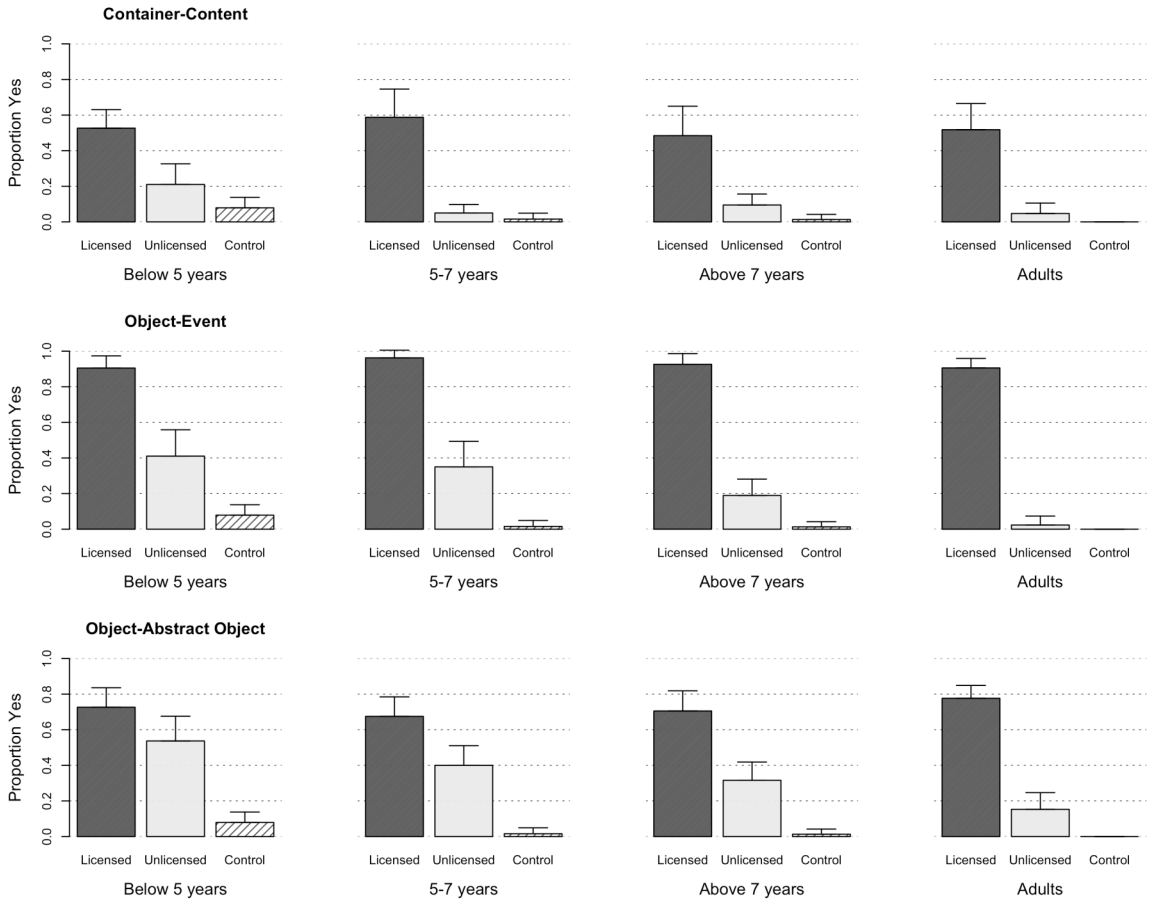

Figure 1: Proportion of "Yes" responses for licensed (dark bars) and unlicensed (light bars) mismatch questions, as well as mismatch control questions (shaded bars) from Section 2.2.2, by age group and shift template. Bars = +2SEM.

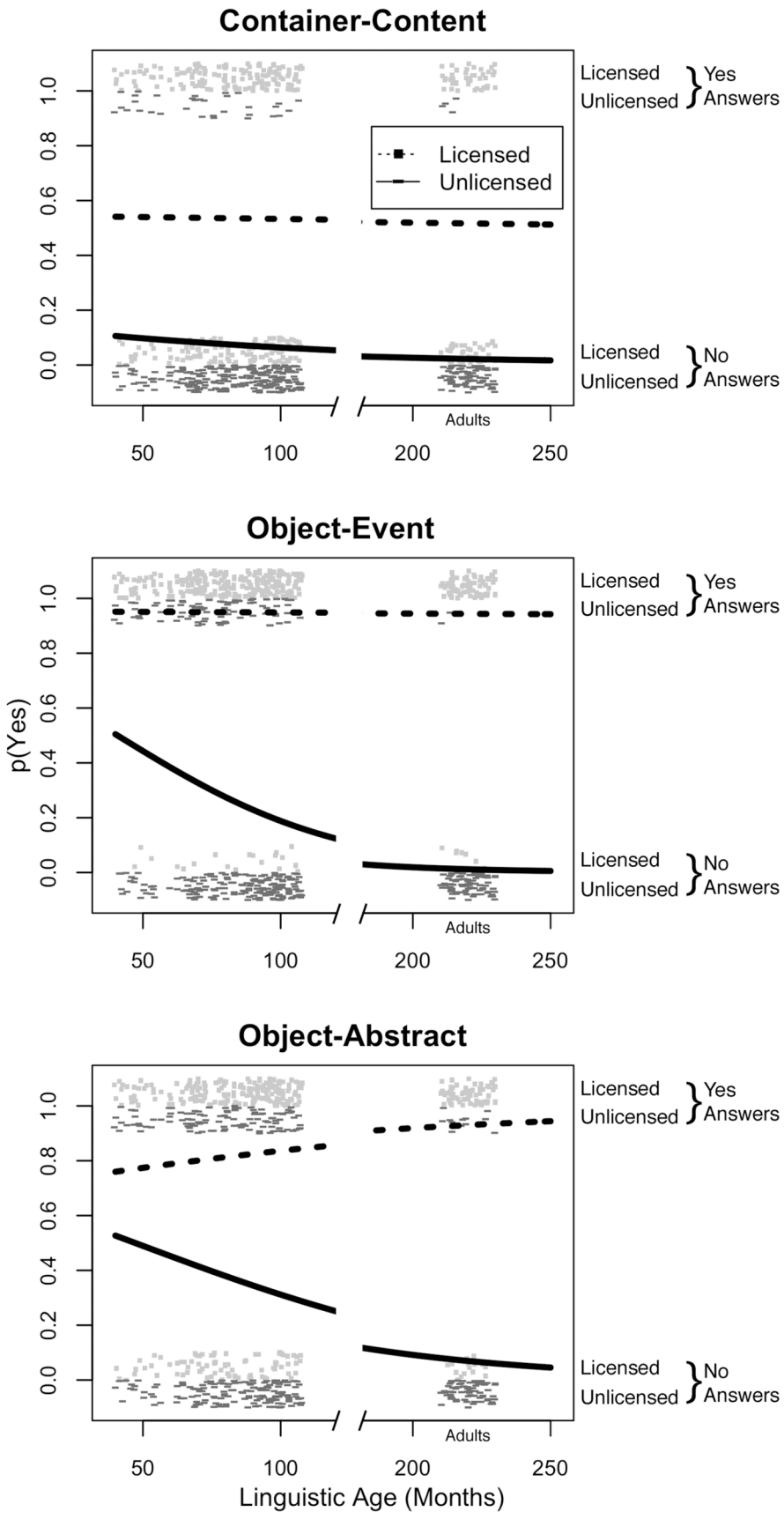
Container-Content
Object-Event
Object-Abstract
p(Yes)
0.0
0.2
0.4
0.6
0.8
1.0
Licensed
Unlicensed
Yes Answers
No Answers
Adults
50
100
200
250
Linguistic Age (Months)

Figure 2: Responses to licensed and unlicensed mismatch questions by age. Individual points represent individual answers of individual subjects. Yes answers are jittered around 1.0, No answers are jittered around 0.0. Answers are segregated by question type, which is also indicated by plotting character (squares = Licensed, dashes = unlicensed). Lines indicate the estimated probability of providing a yes answer to a licensed mismatch question (solid line) and to an unlicensed mismatch question (dotted line) from the models fit in section 2.2.1.

2.2.2 Control analysis I: Animacy violation controls

To test if the high affirmation rates for licensed questions and, for children, unlicensed questions were due to an affirmation bias or to shifting of the relevant sense, we compared their answers to animacy violation control questions. If all affirmations result from a Yes-bias, then the affirmation rate should be similar across all questions.

Again, we used a mixed-effects logistic regression fit to each shift template, with the control condition as our baseline and with language age, two dummy variables coding for a licensed or unlicensed mismatch question, and two interaction terms between language age and the dummy variables, as predictors. We also included random intercepts for subjects and items. Our age predictor was standardized so that other predictors estimated their effects at the youngest age tested, 44 months (see footnote 6).

Figure 3 displays the estimated probabilities of affirming licensed (solid line) or unlicensed (dotted line) mismatch questions over development, as well as the estimated probability of affirming animacy violation controls (dashed line). In addition, the mean proportion of Yes answers by age group and question type can be seen in Figure 1. The

relevant terms of the regressions are described below, while the full regressions are included in Table 2 of the supplementary materials. N for each regression was 994, while deviances were: container-contents = 760, object-event = 569, and object-abstract = 755.

Consistent with the use of shifts in answering mismatch questions, participants affirmed licensed mismatch questions reliably more often than control questions in all three templates (Object-Event: $\beta = 6.47$, s.e. = 0.86, $Z = 7.55$, $p < .01$; Object-Abstract: $\beta = 5.28$ s.e. = 1.42, $Z = 3.70$, $p < .01$; Container-Contents: $\beta = 3.55$, s.e. = 0.81, $Z = 4.37$, $p < .01$). Similarly, unlicensed questions from the object-event and object-abstract templates were affirmed reliably more often than controls (Object-Event: $\beta = 3.24$, s.e. = 0.81, $Z = 3.98$, $p < .01$; Object-Abstract: $\beta = 4.10$, s.e. = 1.42, $Z = 2.90$, $p < .01$). However, the affirmation rate for unlicensed container-contents questions was not reliably different than the rate for controls. ($\beta = 1.34$, s.e. = 0.84, $Z = 1.60$, *ns*).

These results suggest that a response-bias explanation of the affirmation rate is only plausible for the unlicensed container-contents mismatch questions; in the other unlicensed mismatch conditions (object-event and object-abstract), children' high affirmation rates appear to be robust, over and above any possible response bias.

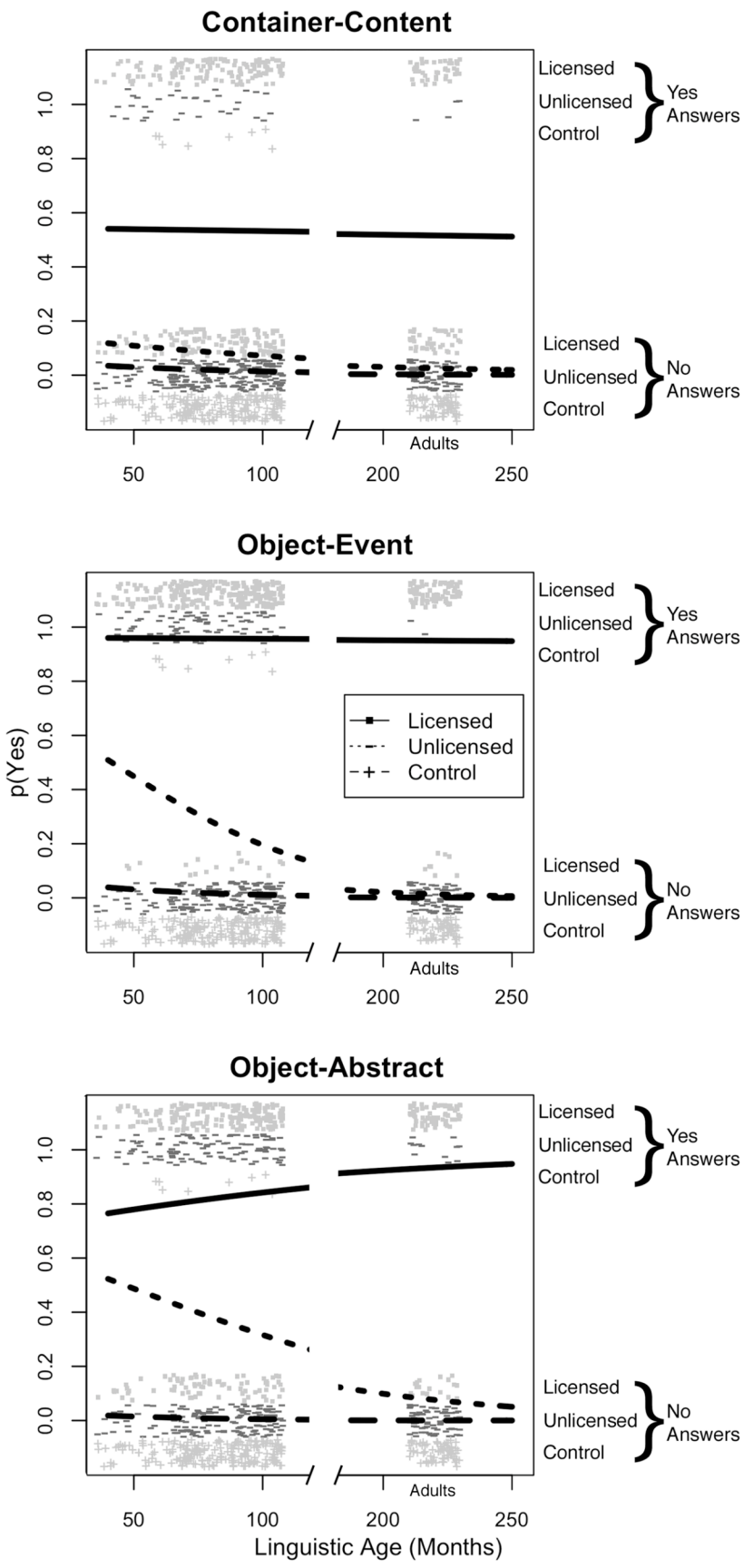
Container-Content
Object-Event
Object-Abstract
p(Yes)
0.0
0.2
0.4
0.6
0.8
1.0
50
100
200
250
Adults
Linguistic Age (Months)
Licensed
Unlicensed
Control
Yes
Answers
No
Answers

Figure 3: Responses to licensed and unlicensed mismatch questions, as well as mismatch control questions, by age. See Figure 2 for full explanation of the plot. Lines indicate the estimated probability of providing a yes answer to a licensed mismatch question (solid line, 'square' character), to an unlicensed mismatch question (dotted line, 'dash' character) and to a control question (dashed line, 'plus' character) from the models fit in 2.2.2.

2.2.3 Control analysis II: *d*'

As a further control for response bias, and to control statistically for some of the vocabulary differences between licensed and unlicensed questions, we assessed changes in children's answers using *d'*, rather than their raw responses. We derived separate licensed and unlicensed *d'* scores per person per shift template. Licensed *d'* compared match questions (e.g. *Could a DVD be round?)* to licensed mismatch questions *(Could a DVD be an hour long?)* and unlicensed *d'* compared match questions (*Could a movie be an hour long?)* to unlicensed mismatch questions (*Could a movie be round?)*. We corrected hit/false alarm rates of 1 to 0.9, and rates of 0 to 0.1 (following Macmillan & Creelman, 2004), based on the number of questions.

Differences between, and changes in, *d'* over age were analyzed using mixed-effects regressions for each shift template, with fixed effects of linguistic age (in months), *d'* type (licensed = 0, unlicensed = 1), and their interaction, as well as a random intercept for subjects. We estimated *p* values using Markov-chain Monte-Carlo simulations of the coefficient estimates (indicated with the nomenclature p_{MCMC}), generated using the pvals.fnc function of the languageR package (Baayen, 2008). Figure 4 displays the

change in licensed *d'* (solid line) and unlicensed *d*' (dotted line) over age. The relevant terms of the regressions are described below, while the full regressions are included in Table 3 of the supplementary materials. N for each regression was 142, while deviances were: container-contents = 294, object-event = 252, and object-abstract = 227. Figure 5 displays *d'* scores by age group for the three shift templates.

Consistent with the results of our previous analyses, there was clear evidence for learning in both the object-event and object-abstract templates (indicated by reliable age by question type interactions, Object-Event: $\beta = 0.005$, s.e. = 0.001, $p_{MCMC} < .01$; Object-Abstract: $\beta = 0.006$, s.e. = 0.001, $p_{MCMC} < .01$), but children appeared to be adult-like in the container-contents template (indicated by a non-significant interaction, $\beta = 0.001$, s.e. = 0.001, *ns*).

Crucially, once vocabulary differences between licensed and unlicensed questions were controlled for by this analysis, even the youngest children responded differently to licensed, compared to unlicensed, mismatch questions, across the templates. Reliable effects of question type indicated that unlicensed *d'* was greater than licensed *d'* across all three templates at 44 months, the youngest age tested (Container-Contents: $\beta = 1.03$, s.e. = 0.14, $p_{MCMC} < .01$; Object-Event: $\beta = 1.41$, s.e. = 0.10, $p_{MCMC} < .01$), although this difference was only slight for the object-abstract template ($\beta = 0.51$, s.e. = 0.11, $p_{MCMC} < .01$), suggesting that even young children are more willing to make shifts that are licensed rather than shifts that are unlicensed.

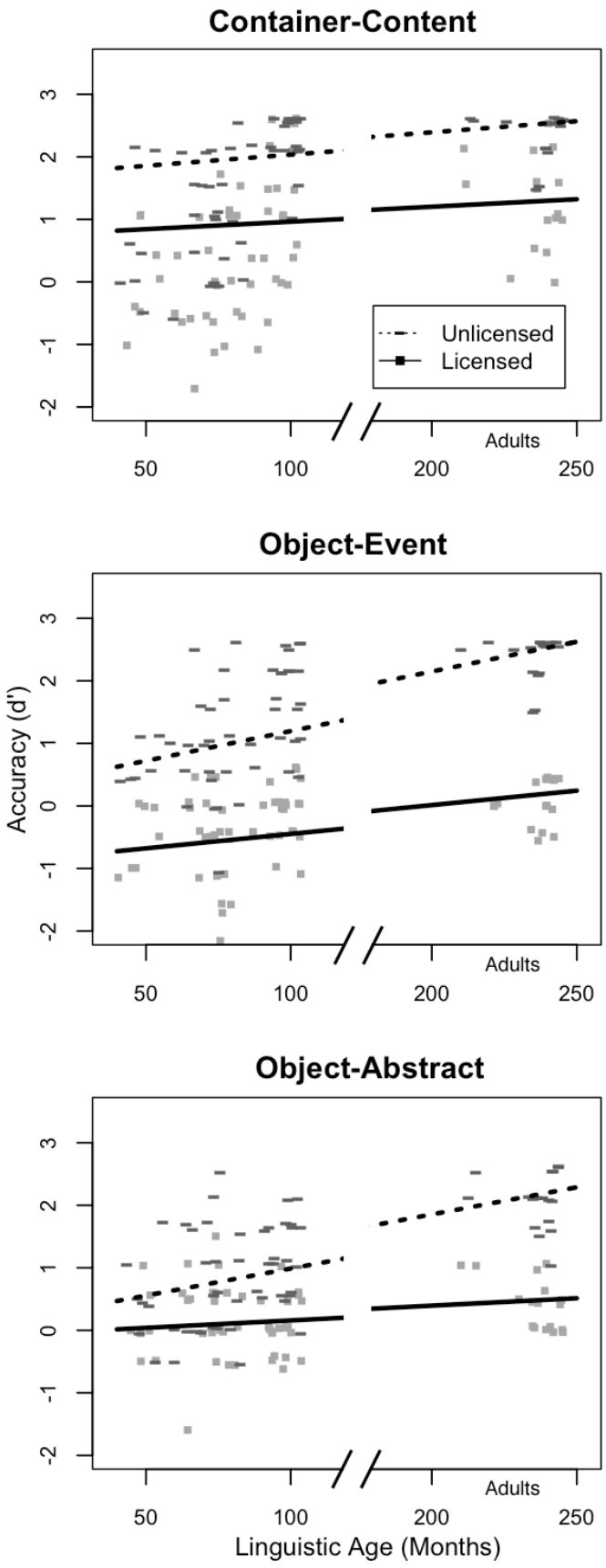
Container-Content
Object-Event
Object-Abstract
Accuracy (d')
Linguistic Age (Months)
Unlicensed
Licensed
Adults
-2
-1
0
1
2
3
50
100
200
250

Figure 4: Change in accuracy (*d'*) over age for licensed and unlicensed questions in the three shift templates. Accuracy refers to participants' ability to accept match questions while rejecting mismatch questions. Lines indicate the estimated *d'* over age for the licensed questions (solid line), and the unlicensed questions (dotted line) from the models fit in 2.2.3.

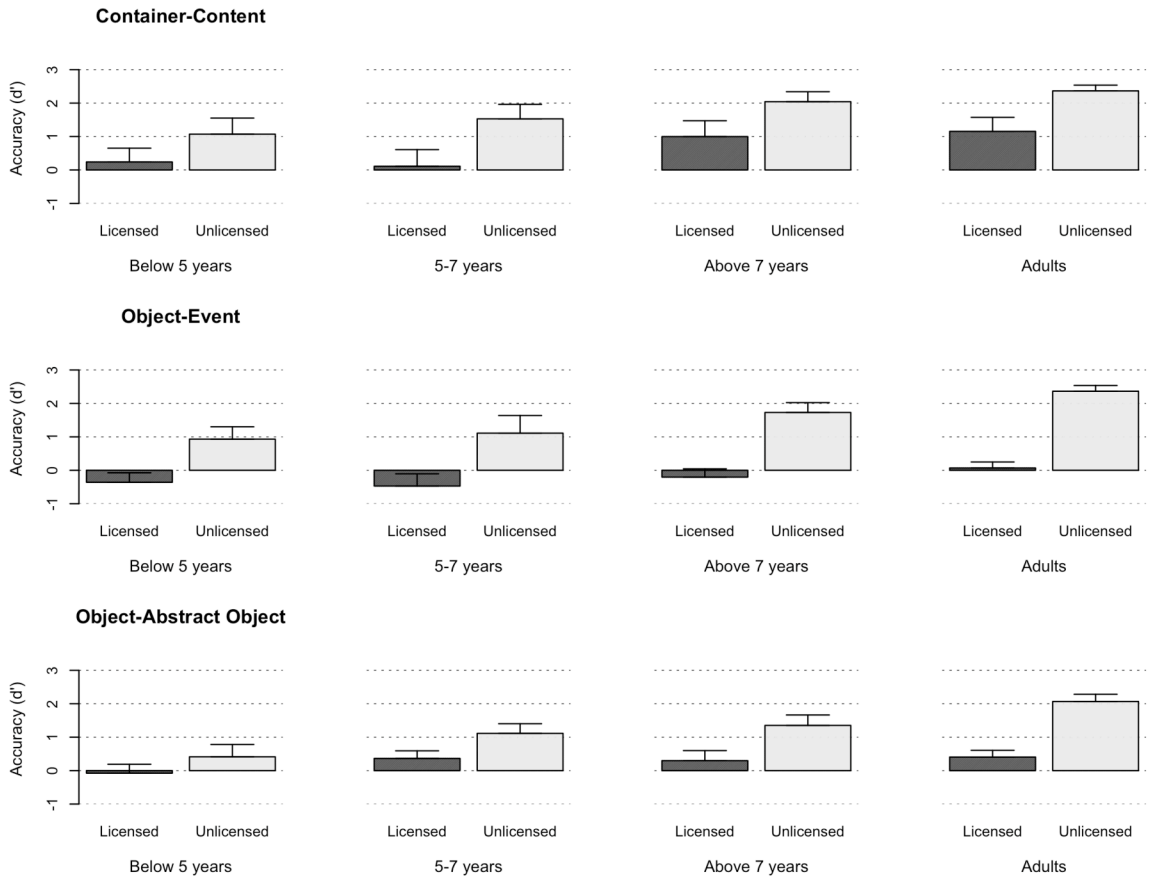

Figure 5: Accuracy (*d'*) by age group and shift template for licensed (dark bars) and unlicensed (light bars) questions. Accuracy refers to participants' ability to accept match questions while rejecting mismatch questions. Error bars = +2 SEM.

2.2.4 Summary of results

Children were adult-like in several ways. They accepted licensed mismatch questions at approximately the same rate as adults, and, for the container-contents template, rejected unlicensed questions at a similar rate. However, in an important respect children were systematically different from adults: young children showed a marked tendency to affirm unlicensed object-event and object abstract questions, above-and-beyond the predicted effects of response bias, and consistent with a productive learning strategy.

2.3 Discussion

Why were children so willing to affirm both licensed and unlicensed mismatch questions? One possibility is that they are productive learners, adult-like in many respects yet profligate in how they assign senses to words. This suggests that at around 5 years-of-age children are wiling to assign *DVD* both a DISC and a FILM sense, as adults do, but are also willing to assign *movie* a DISC sense. With age, they may learn to pare down the set of senses they readily assign to a word, although the point at which this occurs seems to be dependent on the shift; the container-contents shift appeared to be acquired earlier than the object-event or object-abstract shifts. This explanation is relatively parsimonious, in that it does not require children to learn "how" to shift between senses, only to learn exactly which shifts are appropriate. But the problem of how children learn to restrict the set of shifts remains, and we return to this in the discussion.

An alternate interpretation of Experiment 1 is that children may be confused about the meanings of the lexical items under test, believing a DVD to be the sort of thing that

can be an hour long, rather than the sort of thing that can contain an hour-long movie. This idea is not without precedent. Keil (1979) argued that young children suffer from a "collapsed ontology". That is to say, they do not quite understand that ontological sorts such as objects and events are different things with different possible properties. Children could therefore accept mismatch questions without using a shift..

There is also still the possibility of response bias contamination, if response bias varies across domains (see Fritzley & Lee, 2003). That is, if uncertainty results in a Yes-bias, children may be more willing to affirm licensed and unlicensed mismatch questions about DVDs than animate objects because they are less certain about what properties DVDs can take. Again, this predicts that children will accept questions without using a shift, assigning a DISC sense to DVD.

In summary, the over-productivity hypothesis claims that children master shift use at a very early age, but overestimate the set of senses they should accept. The ontological confusion and response bias hypotheses claim that children cannot shift words' senses, but give the appearance of over-productivity because they are unsure which properties apply to which objects. Experiment 2 tests these explanations by asking children to explain their answers. If children can shift between senses, their explanations should differ across the question types. By contrast, if they assume the same sense no matter the predicate, their explanations should be similar across the questions.

Additionally, Experiment 2 assessed whether the learning strategy children took in their shift production was as productive as the one they appear to take in their comprehension. We asked whether children who appear to be willing to comprehend unlicensed shifts were also willing to produce sentences requiring a listener to make the

same shift. Previous work in syntactic development has suggested that children are more constrained in the forms they produce than the forms they can comprehend (e.g., Fernandes et al., 2006; Gertner, Fisher, & Eisengart, 2006; Tomasello, 2000). Is the same true for shift development? Such a result would suggest that children distinguish between shifts that they expect others to make in comprehension, and shifts that they are willing to make in their own comprehension.

3 Experiment 2

Experiment 2 had two aims. First, to contrast the over-productive and ontological confusion/response bias hypotheses. Second, to test whether children follow the same learning strategy in both their shift comprehension and their shift production. The main task was a shortened version of Experiment 1, but each time one of a subset of predicability questions was affirmed, the child engaged in two follow-up tasks.

First, we tested what sense children assigned arguments like *DVD* by asking them to explain their answer. If children are not shifting, and assume that *DVD* refers to DISC in both match and mismatch questions, their explanations for these should be similar, as in both cases they are describing a property that they believe is predicable of a disc. This result would suggest that children actually follow a conservative learning strategy, despite their high affirmation rate. By contrast, if children do follow a productive learning strategy and can shift between senses, they should more frequently use terms related to the shifted MOVIE sense following mismatch questions.

To assess whether children use a conservative or productive strategy in shift production, we next asked them to recall the robot's original question. Productive

children who make both licensed and unlicensed shifts in production should reproduce the question more-or-less accurately. Conservative children should be accurate with match questions and possibly licensed mismatch questions, but modify unlicensed mismatch questions so that they do not require the listener to make an unlicensed shift. In addition, the shift production task acts as an implicit test of children's shift comprehension. A longstanding result in cognitive psychology is that we recall sentences based on a "gist" of their meaning, without recalling the exact sentence structure (Bransford, Barclay, & Franks, 1972; Bransford & Franks, 1971), and so the patterns of production might well reflect children's comprehension. In particular, if children modify lexical items when repeating questions, that would be evidence that they shifted away from the surface interpretation of the Robot's utterance.

The predictions of the hypotheses for the tasks are given in Figure 6. Since shift production is only interesting if children also shift in their comprehension, the hypotheses are laid out accordingly.

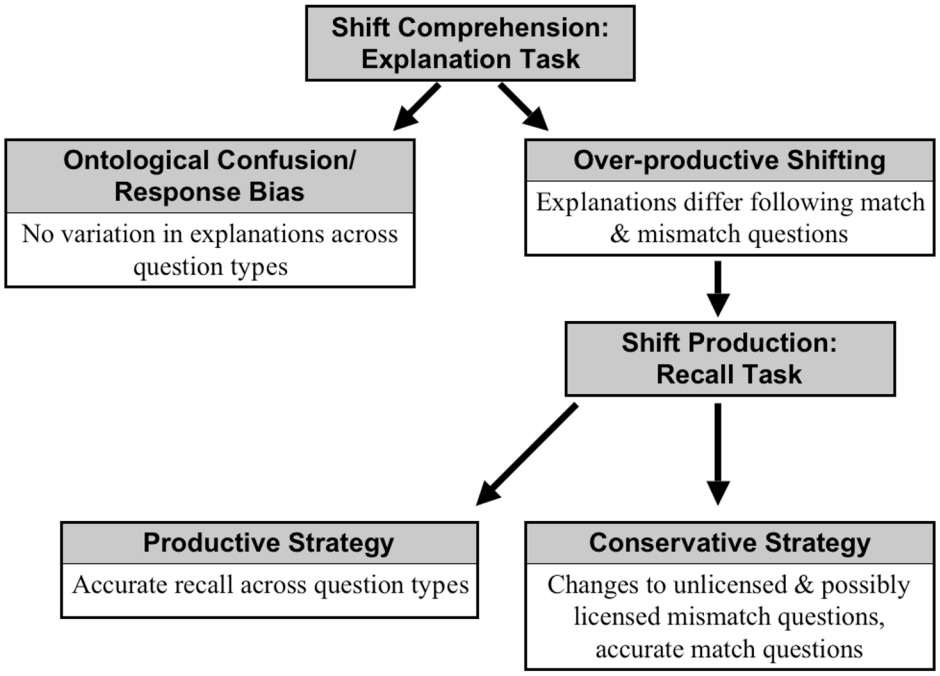

Figure 6: Predictions of the different hypotheses for the follow-up tasks of Experiment 2.

3.1 Method

We used a modified version of the robot task introduced in Experiment 1, with two follow-up tasks.

3.1.1 Participants

20 children aged between 3; 8 and 6; 2 participated. Mean age was 4; 6 (SD = 6 mo). All subjects used English as their first language and had no history of language disorders. Children had a mean linguistic age of 5;7 (SD = 14 mo).

3.1.2 Materials

We used three sets of four predicability questions taken from each of Experiment 1's

three shift templates, picking sets whose match questions had the highest affirmation rate. We did not use a fully specified version of the object-event shift (e.g., *Could a boy begin reading a book?*) as its results were equivalent to the licensed match question in Experiment 1. The questions used are marked in Appendix A.

Presentation order for the questions varied between two lists, with order of mention of licensed and unlicensed shifts counterbalanced over lists and blocks, and a control item separating each block. Follow-up tasks were provided whenever children affirmed a mismatch question, and when they affirmed half of the match questions, with the critical questions pre-specified and counterbalanced between the two ordered lists. The number of follow-ups that could possibly be administered was identical across the question types, because only half the match questions led to follow-ups. In practice, however, children were more likely to affirm Match questions than any other, and consequently these questions led to the largest number of follow-ups.

3.1.3 Procedure

We followed a similar procedure to Experiment 1, with the inclusion of the two follow-up tasks. First, in the Interview task, the experimenter pressed the child to give a reason for her answer. Children were asked how it was possible for the object to have the property, (e.g., *how could a DVD be an hour long?).* Those who wouldn't answer were asked how they could tell whether the object had the property (e.g., *how could you tell if a DVD was an hour long?*), and those who uninformatively repeated that the object had the property were asked what part of the object had the property (e.g., *what part of the DVD is an hour long?*). Figure 7 depicts the interview's structure.

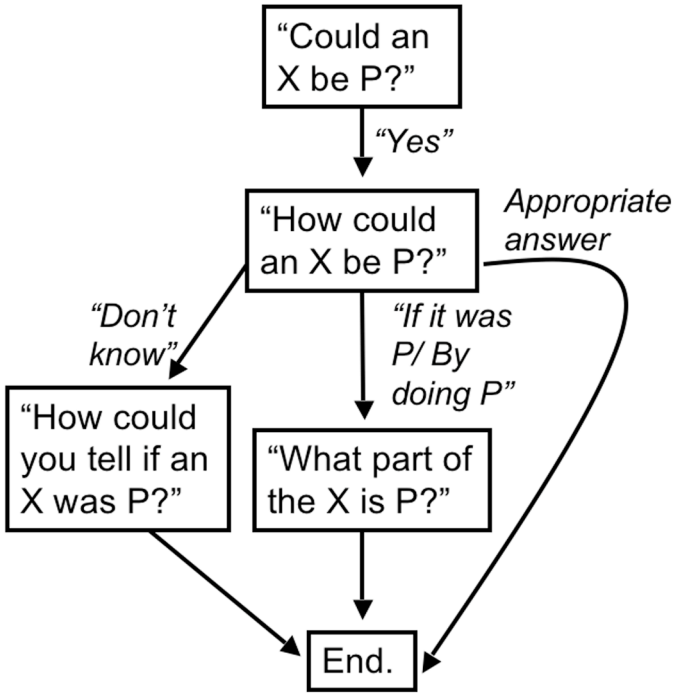

Figure 7: Interview structure for the first follow-up of Experiment 2. Questions are boxed in roman type, answers unboxed in italic type.

To assess production in the recall task, the experimenter asked the participant what question the robot had asked. Rather than have the child repeat back to the experimenter, we used an additional character (a stuffed Winnie the Pooh) that observed the experiment throughout. Before the session the experimenter explained that Pooh often failed to pay enough attention, and that children should listen carefully to what the robot said in order to help Pooh later on. To make the child recall the question, the experimenter told her that Pooh had failed to pay attention during the question, and wanted to find out what the robot had asked. The lag between the robot's original question and the onset of the recall task was not explicitly timed, but was typically 1-2 minutes.

3.2 Analyses and Results

3.2.1: Predicability questions

Children's answers to the predicability questions were recorded and analyzed in a similar manner to Experiment 1, except that regressions did not include fixed effects of age. Tables 4 and 5 of the supplementary materials give full details of the regressions. The results of both a *d'* analysis (All $N = 40$, Deviance: container-contents = 95, object-event = 100, object-abstract = 88) and an analysis of mismatch questions (All $N = 120$, Deviance: container-contents = 119, object-event = 100, object-abstract = 150) were consistent with the results from the younger children in Experiment 1. The mean proportion correct, and the mean *d'* for each condition are displayed in Figure 8. First, there were only marginal differences between children's answers to licensed and unlicensed object-abstract questions (*d'*: Mean Licensed = 0.38 (SD = 0.76), M Unlicensed = 0.72 (SD = 0.80), $\beta = 0.34$, s.e. = 0.19, $p_{MCMC} = .16$; Mismatch questions: M Licensed = 0.70 (SD = 0.28), M Unlicensed = 0.57 (SD = 0.34), $\beta = -0.69$, s.e. = 0.40, $Z = 1.71$, $p = 0.09$). Second, both analyses indicated that licensed mismatch questions were affirmed reliably more often than unlicensed questions in the object-event and container-contents shifts, suggesting that children do distinguish between licensed and unlicensed shifts in these conditions (even if they still make some unlicensed shifts) (Object-Event: *d'*: M Licensed = -0.68 (SD = 0.76), M Unlicensed = 0.21 (SD = 1.14), $\beta = 0.89$, s.e. = 0.19, $p_{MCMC} < .01$; Mismatch questions: M Licensed = 0.93 (SD = 0.14), M Unlicensed = 0.58 (SD = 0.36), $\beta = 3.39$, s.e. = 1.16, $Z = 2.93$, $p < .01$; Container-Contents *d'*: M Licensed = 0.086 (SD = 0.96), M Unlicensed = 1.33 (SD = 0.76), $\beta =$

1.24, s.e. = 0.20, p_{MCMC} < .01; Mismatch questions: M Licensed = 0.58 (SD = 0.34), M Unlicensed = 0.10 (SD = 0.16), β = -2.76, s.e. = 0.53, Z = 5.21, p < .01), Additionally, participants were still highly likely to affirm unlicensed object-event shifts, although less likely than they were to affirm licensed object-event shifts. Overall, these results were broadly similar to those in Experiment 1: Young children were willing to accept all licensed mismatch questions, and accepted a range of unlicensed mismatch questions as well.

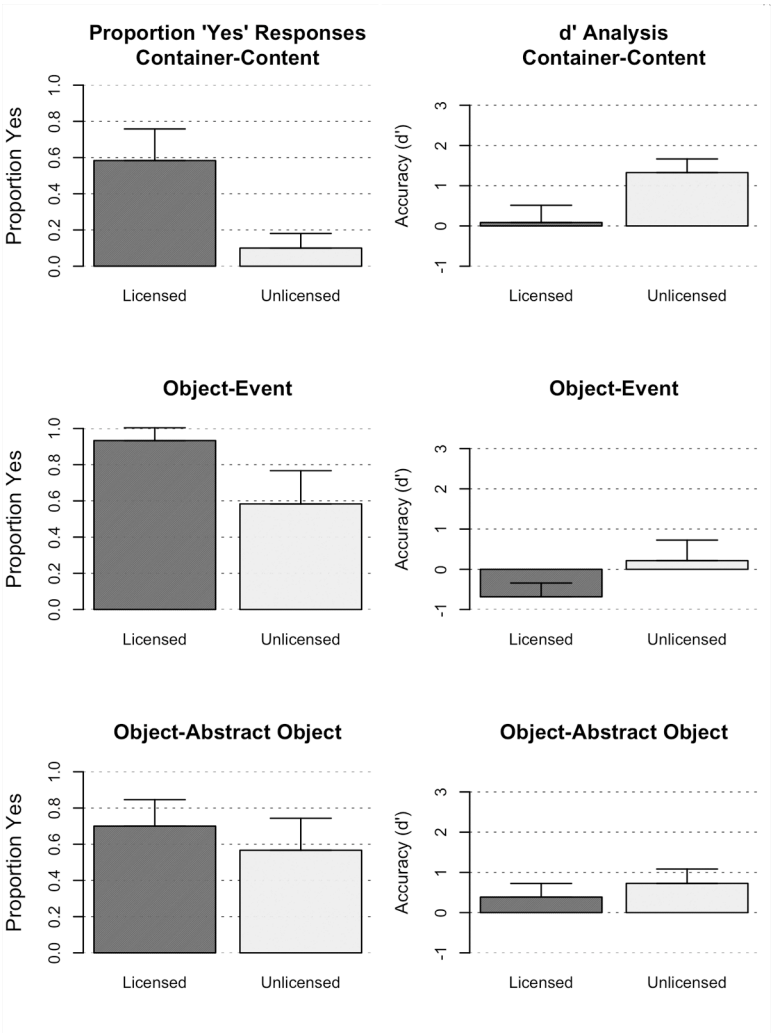

Figure 8: Mean proportion 'Yes' responses and mean *d'* score for each condition in the

predicability question section of Experiment 2. Dark bars represent results for licensed conditions, and light bars for unlicensed. Error bars = +2 SEM.

3.2.2: Follow-up questions

When analyzing the follow-up tasks, we compared answers following licensed and unlicensed mismatch questions to those following match questions (not differentiating between licensed and unlicensed match questions), collapsed across the shift templates to maintain an adequate sample size. Because follow-up questions were only administered following affirmations, the number of answers provided differed across subjects and questions, resulting in an unbalanced dataset. To reduce concerns about the imbalance, all of our regressions included the factors that we believed determined whether data was missing: individual subject intercepts, and predictors coding for the different predicability questions. The data can therefore be considered missing at random.

To analyze the data, two raters judged each answer from transcriptions of the interviews. Criteria for judgments are set out below, as are Cohen's κ's to quantify inter-rater agreement. Items on which they disagreed were excluded from analyses.

3.2.2.1 Interview Task

The interview task assessed why children affirmed the robot's question. If children are ontologically confused, then they should provide similar explanations following both match and mismatch questions, as the same sense is assigned in each. If they are shifting on mismatch questions, their explanations following these should differ from explanations following match questions. We attempted to code explanations in the

simplest manner possible; our resulting scheme classified them into one of three forms. "NA" explanations contained no relevant information. "Shifted" explanations contained wording indicative of the shifted sense of the critical argument. For example, if children were asked *How could a DVD be an hour long?,* or *How could a movie be round?* or *How could a boy could start a picture?* and their answer referenced *watching, DVDs* or *drawing*, respectively, this would be counted as a Shifted explanation (examples include '"*How could a boy start a picture?" "He draws it"'* and '"*How could a movie be round*?" "*If the DVD was round")*. All other answers were classified as "Unshifted" explanations. This group included answers that did not reference the shifted argument (e.g., "How could a movie be an hour long?" "Because sometimes movies are really long*"*) as well as answers that did not specify the sense of the argument (e.g., "How could a cup be spilled?" "I [sic] might fall on the floor*"*). Such answers were collapsed to minimize subjective judgments during coding. Inter-rater agreement was substantial, as measured by a Cohen's κ of 0.77.

We tested whether the probability of providing either a shifted explanation or an unshifted explanation changed between question types using a mixed effects logistic regression. Answers to the match questions were used as baseline, and two dummy variables coded whether a question was a licensed mismatch, or an unlicensed mismatch. We also included random intercepts for subjects and items (number of observations = 205, deviance = 190). Full details of the regression are given in Table 6 of the supplementary materials.

If children perform shifts on mismatch questions but not match questions, then they should produce more shifted explanations following the former. Otherwise, they should

not. The total number of each explanation type, by question type (matching, licensed shift, unlicensed shift), is displayed in Figure 9.

When explaining the answer to Match questions, children produced reliably more unshifted than shifted explanations (Mean no. of Unshifted explanations = 4.40 (SD = 1.63), $\underline{M}$ Shifted = 0.30 (SD = 0.47), , β = -2.91, s.e. = 0.59, Z = 4.94, p <.01), but the proportion of shifted explanations reliably increased when the question involved a licensed or unlicensed shift (Licensed: Unshifted $\underline{M}$ = 1.75 (SD = 1.5), Shifted $\underline{M}$ = 2.85 (SD = 1.22), β = 3.62, s.e. = 0.66, Z = 5.44, p < .01; Unlicensed: Unshifted $\underline{M}$ = 0.88 (SD = 1.05), Shifted $\underline{M}$ = 1.94 (SD = 1.83), β = 4.20, s.e. = 0.74, Z = 5.71, p <.01). This increase in proportion was similar across the two mismatch conditions. Children's explanations were therefore consistent with their use of a shift to interpret both the licensed and unlicensed mismatch questions, as predicted by the over-productive shifting hypothesis.

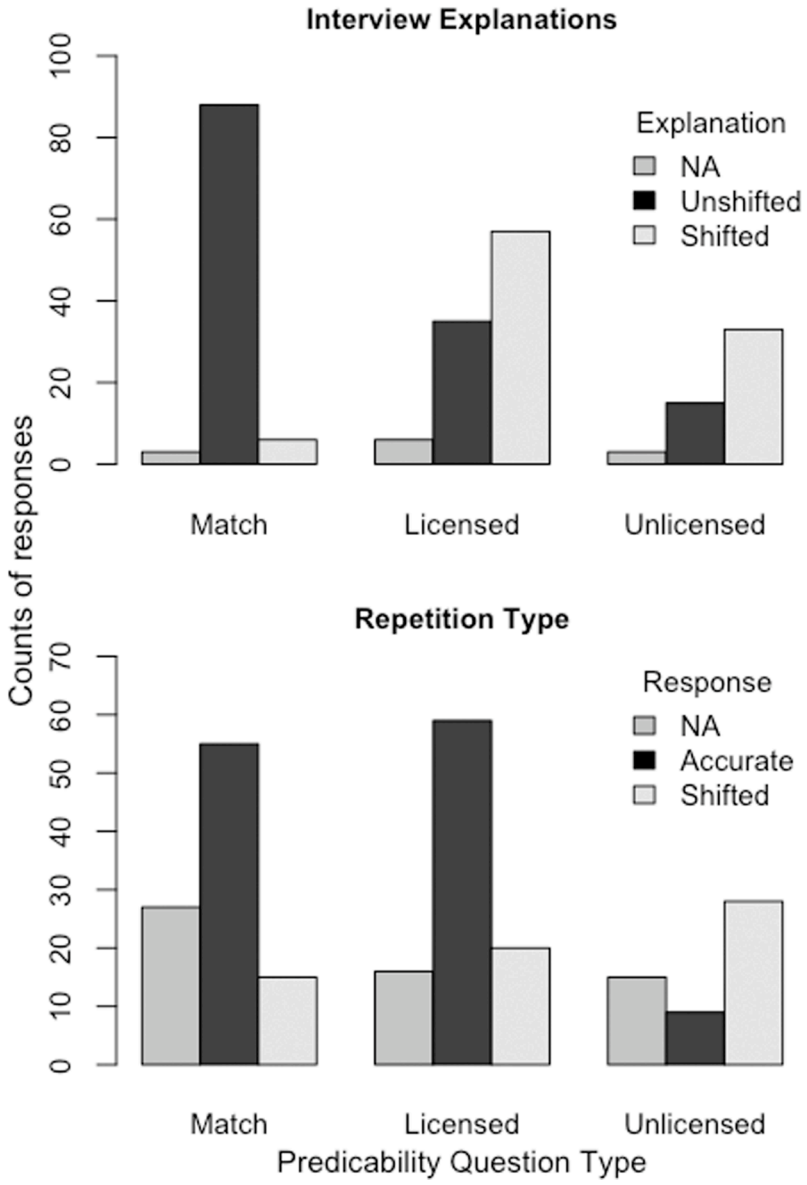

Figure 9: Counts of response types across the follow-up tasks of Experiment 2, by predicability question type. Top graph: Interview task. Counts of the three different explanation types coded following match questions, licensed mismatch questions and unlicensed mismatch questions. Bottom graph: Recall task. Counts of the three different repetition types coded following the three question types.

3.2.2.2 Recall and Production Task

To analyze production, children's recall of the Robot's question (e.g. *Could a song be shiny*?) was classified as one of three types. "Accurate" repetitions (e.g., *Could a song be shiny*?) closely matched the robot's question. "Shifted" repetitions used an alternative word that referred to the shifted sense of the question's argument (e.g., *Could a CD be shiny*?). "NA" repetitions did not refer to either meaning. Some leeway was given in coding; the child's recall did not have to be completely accurate (for example, *The song was shiny* would be coded as an accurate repetition of *Could a song be shiny?*). The same raters as before judged each answer, excluding items on which they disagreed. Agreement was again substantial, Cohen's $\kappa = 0.81$. To compare the probability of providing an accurate or shifted response, we fit a multi-level logistic regression as in the interview analysis (number of observations = 186, deviance = 194). Full details of the regression are given in Table 6 of the supplementary materials.

If children are productive, their recall should be similarly accurate across all conditions. Conservative children should willingly repeat match questions, and possibly some licensed mismatch questions, but change the wording of unlicensed mismatch questions and unfamiliar licensed mismatch questions in order not to make an unlicensed shift. Figure 9 displays the results.

Children typically provided an accurate recall for match questions (Accurate $\underline{M}$ = 2.75 (SD = 2.0), Shifted $\underline{M}$ = 0.75 (SD = 0.97), $\beta = -1.49$, s.e. = 0.40, $Z = 3.68$, $p < .01$) and there was no reliable difference in their accuracy following a licensed mismatch question (Accurate $\underline{M}$ = 2.95 (SD = 1.89), Shifted $\underline{M}$ = 1.00 (SD = 1.17) $\beta = 0.16$, s.e. =

0.42,Z = 0.37, *ns*). However, children were less likely to give an accurate recall following unlicensed mismatch questions and instead were reliably more likely to provide a shifted recall (Accurate $\underline{M}$ = 0.47 (SD = 0.70), Shifted $\underline{M}$ = 1.47 (SD = 1.22), β = 2.10, s.e. = 0.42, p <.01), changing the referent of a sentence to accord with the shift we predicted they had made in their comprehension[7].

This suggests that children follow a conservative learning strategy in their production, even while using a productive strategy in their comprehension. They willingly comprehend both licensed and unlicensed shifts and, when asked to do so, will produce sentences whose words require the listener to make a licensed shift. But when prompted to use a word in a sense that is not licensed they refuse, and instead change the lexical item.

This modification toward the shifted meaning additionally provides converging evidence that children interpret unlicensed mismatch questions using a shift. People typically remember sentences in terms of their gist; children's production suggests that their gist was consistent with a shifted meaning.

3.3 Summary of results and discussion

Experiment 2 replicated a main result of Experiment 1: young children accept predicability questions that are only interpretable following a licensed shift, and in addition accept questions requiring an unlicensed shift.

[7] One concern is that the change in repetition type might be caused by the NP's complexity in the Unlicensed Object-Event MisMatch questions (e.g., *Could drawing a picture be large?*). However, the paired NP in Match questions is also complex (*Could drawing a picture be quick?*), and there is still a reliable increase in the probability of shifting in the Unlicensed condition (p < 0.01). Furthermore, the probability of changing the Match question does not vary across the different shift templates (all p > 0.15).

Furthermore, it provided evidence that, in comprehension, children accept a systematically broader range of senses than adults typically license, and hence that Experiment 1's results are not due to ontological confusion or response bias. Specifically, children gave different explanations for their answers following mismatch questions, compared to match questions. In particular, they more often invoked explanations involving the shifted sense following mismatch questions, consistent with the ability to shift between multiple senses of a word during normal language comprehension.

Finally, a recall task indicated that children use a conservative learning strategy in the online generation of shifts. Our participants accurately repeated match predicability questions and licensed mismatch questions, but were more likely to reconstruct unlicensed mismatch questions as a licensed paraphrase, a result that also favors the interpretation that children overestimate which senses can be assigned to words during comprehension.

4 General Discussion

When we talk about the world around us, we try to speak coherently and sensibly. For example, upon listening to a newly purchased album, we might well describe the first song as beautiful, or loud, but would hesitate to proclaim it shiny. However, human language frequently ignores the dicta of sensicality and category. A CD can be loud, or a DVD an hour long, despite being inert pieces of plastic. How do children learn which of an infinite number of possible senses are licensed? The results of our two experiments indicate that, in comprehension, children follow a productive strategy to accomplish this feat. They were adult-like in many respects, making licensed shifts without difficulty, but

they also over-generalized by accepting a range of shifts that are unlicensed for adults. For a young child, hearing a movie described as round seemed almost as interpretable as hearing a DVD described as an hour long. Over development, children narrow down the set of shifts that they are willing to make. The extended period over which this development occurred—children were not fully adult-like until around 7-years—contrasts with a long literature on the rapidity of children's word learning. It suggests that although children may quickly map the initial referent of a word, the more subtle components of lexical semantic development take place over a much longer time period.

Although young children were willing to accept unlicensed shifts in comprehension, their production strategy appeared conservative. Children accurately repeated questions containing licensed shifts, such as *Could a DVD be an hour long?* but balked at the prospect of using an unlicensed shift when repeating questions like *Could a movie be round?*. Rather, this situation drove children to search for an alternative wording, altering the robot's question while retaining the meaning derived from their initial unlicensed shift (e.g. assigning *Movie* a DISC sense).

This dissociation between comprehension and production is not entirely surprising. Using a productive comprehension strategy provides advantages (such as understanding the shifts of those around you), as well as disadvantages (such as overestimating the shifts of those around you). However there is no obvious benefit to producing a previously unencountered shift; it only makes misunderstandings more probable. The only situation in which children might need to produce an unencountered shift is when they have a lexical gap: for instance, if they lacked the word *DVD* they may use a shifted meaning of *Movie*. A decrease in the number of lexical gaps has already

been proposed as an explanation of children's reduced use of class extensions over development (e.g., Can you lipstick the trashcan?, Bushnell & Maratsos, 1984, Clark, 1981), so it may be that shift production would be less conservative amongst younger children with smaller lexicons.

Nevertheless, there is a certain paradoxical quality to the asymmetry: Children know which shifts they can and cannot expect others to comprehend, but are much looser in the shifts they comprehend themselves. This pattern is consistent with a broader literature in which children limit their production to previously heard forms, while allowing greater generalization in comprehension. For example, 2- to 3-year-old children typically only produce verbs inside the argument structures that they have heard demonstrated in the input (Tomasello, 1992, 2000) but are nonetheless able to generalize these verbs to novel argument structures in comprehension (Fernandes et al., 2006; Gertner et al., 2006). Our results suggest that children restrict their lexical semantics in an analogous way.

One as-yet-unexplained result emerged from these experiments: Both children and adults were less willing to make licensed container-contents shifts, relative to object-event and object-abstract shifts, and the container-contents template became adult-like earliest. One possibility is that both these differences can be explained by the syntactic context in which the to-be-shifted argument was encountered. The argument of the licensed mismatch questions (*Could a pot be stirred?*) carried an indefinite article, which may have cued an object sense, conflicting with the cue from the predicate, which demanded that the sense be a substance. Likewise, the arguments of the unlicensed container-contents mismatch questions (*Could some soup be cracked?*) carried *some* as a

determiner, which cues a substance sense, conflicting with the object demanded by predicates like *cracked*. These clashes may have led fewer participants to make each shift, resulting in the lower acceptance rates.

However, the lower overall acceptance rate is not the only explanation of the container-contents shift's early acquisition. In many theories of adult shift use, the set of licensed shifts is partially determined by background knowledge relating the different possible senses. In addition, studies of metaphor development suggest that knowledge of both source and target domain is vital for early comprehension (Keil, 1986), so it would not be too surprising if children's shift use were similarly dependent on their background world knowledge. Given that children appear to know a lot more about physical objects than abstract objects or events (Keil, 1979), the rapid acquisition of the container-contents shift (but neither other shift template tested) is explicable, as both senses describe physical objects.

Children's willingness to make shifts during comprehension is consistent with previous work on children's comprehension of metaphor (Vosniadou, 1987), generically quantified sentences (Gelman & Bloom, 2007), mass-count ambiguity (Barner & Snedeker, 2005), and class extensions (Bushnell & Maratsos, 1984), all of which support an early-emerging understanding of complex lexical semantic phenomena. The present work goes further in showing that children can perform such semantic operations without the syntactic cues to meaning provided for comprehension of generics. This is particularly impressive because learning to use non-syntactic cues for sense resolution is an extremely difficult problem, which has posed major challenges for computer scientists working in natural language processing. Children's success leaves an open question as to

whether there are information sources they use in sense resolution that natural language processing algorithms cannot.

However the productive strategy here lies in contrast to previous evidence that children fail to compute scalar implicatures, and therefore are wedded to the surface truth conditions of a sentence (Noveck, 2001). This is in some ways surprising, as, for both phenomena, children have to use the context of a constituent to go beyond its surface interpretation. One possibility is that the different pattern emerges because the two phenomena occur at distinct representational levels: shifts are a lexical semantic phenomenon, while scalar implicatures are pragmatic. However, a mechanistic account of why they differ will require a more detailed understanding of the processing steps of each phenomenon.

What develops in shift development?

Most crucially, these experiments indicate that young children are perfectly capable of performing an operation as complex as a shift. Their difficulty appears to lie in understanding the limits of its application, not in recognizing the very possibility of shifting. But how do children eventually come to have an adult-like understanding of the scope and limits of shifts?

This depends on what, exactly, is developing. One interpretation of the data, indeed the one we have taken so far, is that children re-weight the probabilities of different senses over development, or begin to use context differently. For example, although children may initially assign a high probability to a DISC sense for *Movie* (by,

for example, assuming an abstract object-object rule) that probability may decrease over development. Consequently children would use that sense less.

An alternative explanation is that children's changing answers reflect a metalinguistic advance in judging whether a sense is truly appropriate. That is to say, young children compute senses in the same way as older children and adults, but they also have a lower criterion for saying that a sense is appropriate. With development, that criterion shifts until children only accept adult-like meanings. This sort of pattern is evidenced in children's processing of class extensions (Bushnell & Maratsos, 1984): children can interpret these reasonably accurately at around 2 years, but will not judge them to be linguistically anomalous until 7 years.

Most plausibly, both of these factors are operating during shift development. But there are reasons to suspect that the major developmental change is in how meanings are assigned, rather than children's judgments of anomaly, even though we lacked a measure of children's metalinguistic judgments. First, our sentences never actually contained any structural anomalies and were always readily interpretable, so the metalinguistic component of our judgments was probably lower than for Bushnell & Maratsos (1984). Second, we saw no evidence for a developmental change for the licensed container-contents shifts. In this condition, adults accepted the questions only 50% of the time. If children have initial metalinguistic difficulty, we might expect the youngest children to accept these questions more often than adults, but in fact the rate was constant across ages.

What, then, might cause developmental change in how meanings are assigned? One possibility is that young children initially pay *too* much attention to context. The

process of situational fit that we have suggested children use—building a partial representation of the meaning of a phrase and determining the word sense which best fits that meaning and the external situation— clearly provides an extremely powerful tool for assigning a meaning to a word. Presumably, situational fit is most useful during the early stages of language acquisition, when children are unsure of both word meanings and the rules of syntactic/semantic composition. But as an overall strategy, its utility should decline as children learn the rules of their language, and so it should be deemphasized. It may be that younger children's greater willingness to make unlicensed shifts derives from their greater reliance on situational fit.

But if profligate shifting is only caused by an over-reliance on context, this profligacy should be demonstrated across the board, and not just in a subset of the templates. Context, then, can only be part of the developmental story. This suggests a second component: children have to determine which senses are more plausible independent of context, that is to say, figuring out that *DVD* clearly has a FILM sense, but that *movie* is not usually assigned a DISC sense. Exactly how this could occur is unclear: trimming the size of a productive system presents a clear problem for children, in the absence of any negative evidence.

The type of mechanism used to reduce the set of licensed shifts will clearly depend on the representational status of shifts in adults, a subject of some controversy. As discussed in the Introduction, the dominant theories of meaning shifts treat them as a set of rules, each of which acts on words falling into a semantic class, and predictably changes their meanings (as in an object-abstract object rule that transforms DVD's to their contents). One possibility is that children initially possess a broad range of lexical

rules (like Object-Event and Event-Object) that produce both licensed and unlicensed senses (following Copestake & Briscoe, 1995; Dolling, 1995). Over time, they then "unlearn" the rules that produce senses that are unlicensed in their language. In part, this could be accomplished by ruling out senses that are only used infrequently, a strategy (known as entrenchment) that has been proposed to aid in the acquisition of verb argument structure (Ambridge, Pine, Rowland, & Young, 2008; Braine & Brooks, 1995). For example, if children never hear *movie* used to refer to a solid object they might discount the probability that an abstract object-object rule is used in the language.

Unlearning could be aided by innate constraints on possible rules, in a parallel to theories of argument structure acquisition. For example, Pinker et al. (1987) propose that children treat different subsets of verbs as *a priori* less likely to passivize. They are very willing to passivize canonical action verbs, with an agent as subject and theme as object, reliably less willing to passivize noncanonical action verbs, which take themes as subjects and agents as objects, and adopt an intermediate stance toward nonactional verbs (e.g., verbs of perception). These constraints map onto cross-linguistic variation: canonical action verbs are typically passivizable, noncanonical action verbs are not, while the passivization of nonactional verbs varies both across languages, and across semantic classes within a language.

The current paper provides some tentative evidence consistent with constraints on shifting: even the youngest children tested were consistently more likely to affirm licensed questions than unlicensed. But it is certainly not clear that this was the result of innate restrictions, rather than prior learning. Future work on this topic will require a theory of why some shifting rules might be more plausible than others, as well as

corroborating evidence on the distribution of shifts across languages. As it is now, we have few theories as to why certain shifts appear more plausible.

The most prominent theories of shift restrictions propose that the plausibility of a sense is determined by a particular structure in our conceptual organization. Pustejovsky (1995) has proposed a model in which each word is associated with a 'qualia structure', a listing of explanatory modes under which an individual concept can be construed, and these qualia license particular senses. Nunberg (1979) argues that high cue validity between two senses (that is, the predictability of sense A's referent in the presence of B's referent) will make a sense more plausible.

Nunberg's theory is particularly interesting, because it suggests that apparent rules may instead be an artifact of our conceptual organization. For example, we may appear to possess a container-contents rule because, for every container, it is very predictable that it will have contents, and therefore we can use that sense. By contrast, there are an almost unlimited number of things that can be contained, and most of those things do not have predictable containers, so there should be few contents-container shifts in a language. This means that shift use should not be not all-or-nothing, which correctly predicts the acceptability of certain phrases that do not fit into any rules, such as *the unsuccessful movie took up a lot of shelf-space.* The theory also has a major consequence for development: Because there are no linguistic rules to acquire, the main learning challenge for children will be determining which components of conceptual structure are critical for shift use. Once this is established, accurately determining the senses available for each word should arise automatically, as a byproduct of acquiring an adult-like conceptual organization.

Acknowledgments

Many thanks to Gregory Murphy for extremely helpful comments and discussion at many different stages of this project. Thanks also to undergraduate research assistants Tracie Lin and Hanna Gelfand. This work was supported by NIH research grant R01-HD48733.

References

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

Baayen, R. H. (2008). languageR: Data sets and functions with Analyzing Linguistic Data: A practical introduction to statistics. *R package version 0.953*.

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

Barner, D., & Snedeker, J. (2005). Quantity judgments and individuation: evidence that mass nouns count. *Cognition, 97*, 41-66.

Bates, D., & Sarkar, D. (2008). lme4: Linear mixed-effects models using S4 classes.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, Mass.: MIT Press.

Bowerman, M. (1987). Commentary: Mechanisms of Language Acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 443–466). Hillsdale, NJ: LEA.

Braine, M. (1971). On two types of models of the internalization of grammars. In D. Slobin (Ed.), *The ontogenesis of grammar* (pp. 153-168). New York, NY: Academic Press.

Braine, M., & Brooks, P. (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In M. Tomasello & W. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs*. Hillsdale, NJ: LEA.

Brandone, A. C., & Gelman, S. A. (2009). Differences in preschoolers' and adults' use of generics about novel animals and artifacts: A window onto a conceptual divide. *Cognition, 110*, 1-22.

Bransford, J., Barclay, J., & Franks, J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology, 3*, 193-209.

Bransford, J., & Franks, J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2*, 331-350.

Brennan, J., & Pylkkänen, L. (2008). Processing events: Behavioral and neuromagnetic correlates of aspectual coercion. *Brain and Language, 106*, 132-143.

Bushnell, E. W., & Maratsos, M. P. (1984). "Spooning" and "Basketing": Children's Dealing with Accidental Gaps in the Lexicon. *Child Development, 55*, 893-902.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and reports on child language development, 15*, 17-29.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*, 1032-1053.

Cimpian, A., & Markman, E. M. (2008). Preschool children's use of cues to generic meaning. *Cognition, 107*, 19-53.

Clark, E. V. (1981). Lexical innovations: How children learn to create new words. In W. Deutsch (Ed.), *The child's construction of language* (pp. 299-328). London: Academic Press.

Clark, E. V. (1982). The young word maker: A case study of innovation in the child's lexicon. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 396). New York: CUP.

Copestake, A., & Briscoe, E. J. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics, 12*, 15-67.

Crain, S., & Thornton, R. (2000). *Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics*. Cambridge, MA: MIT Press.

Dolling, J. (1995). Ontological domains, semantic sorts and systematic ambiguity. *International Journal of Human-Computer Studies, 43*, 785-807.

Fernandes, K. J., Marcus, G. F., Di Nubila, J. A., & Vouloumanos, A. (2006). From semantics to syntax and back again: argument structure in the third year of life. *Cognition, 100*, B10-20.

Fritzley, V. H., & Lee, K. (2003). Do young children always say yes to yes-no questions? A metadevelopmental study of the affirmation bias. *Child Development, 74*, 1297-1313.

Gelman, S. A., & Bloom, P. (2007). Developmental changes in the understanding of generics. *Cognition, 105*, 166-183.

Gelman, S. A., Goetz, P. J., Sarnecka, B. W., & Flukes, J. (2008). Generic language in parent-child conversations. *Language Learning and Development, 4*, 1-31.

Gelman, S. A., & Raman, L. (2007). This cat has nine lives? Children's memory for genericity in language. *Developmental Psychology, 43*, 1256-1268.

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

Gordon, P. (1996). The Truth Value Judgment Task. In D. McDaniel, C. McKee & H. Smith Cairns (Eds.), *Methods for Assessing Children's Syntax* (pp. 211-231). Cambridge, MA: MIT Press.

Harris, J., Pylkkänen, L., McElree, B., & Frisson, S. (2008). The cost of question concealment: eye-tracking and MEG evidence. *Brain and Language, 107*, 44-61.

Hresko, W. P., Reid, D. K., & Hammill, D. D. (1999). The Test of Early Language Development 3. Austin, TX: Pro-Ed.

Jackendoff, R. S. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*434-446.
Kamei, S.-i., & Wakao, T. (1992). *Metonymy: reassessment, survey of acceptability, and its treatment in a machine translation system*. Paper presented at the Proceedings of the 30th annual meeting of the Association for Computational Linguistics.
Keil, F. C. (1979). *Semantic and Conceptual Development: An Ontological Perspective*. Cambridge, MA: HUP.
Keil, F. C. (1986). Conceptual domains and the acquisition of metaphor. *Cognitive Development, 1*, 73-96.
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: UCP.
Lapata, M., & Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics, 29*, 261-315.
Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, Il: University of Chicago Press.
MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40.
Macmillan, N., & Creelman, C. (2004). *Detection Theory: A User's Guide*. Hillsdale, NJ: LEA.
Manning, C. D., & Schütze, H. (2002). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
McCarthy, D., Koeling, R., Weeds, J., & Carroll, J. (2004). *Finding predominant word senses in untagged text.* Paper presented at the Proceedings of the 42nd annual meeting of the association for computational linguistics, Barcelona, Spain.
McElree, B., Traxler, M. J., Pickering, M. J., Seely, R. E., & Jackendoff, R. (2001). Reading time evidence for enriched composition. *Cognition, 78*, B17-25.
Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*, 39-41.
Miller, G. (1999). On knowing a word. *Annual Reviews in Psychology, 50*, 1-19.
Murphy, G. L. (2001). Fast-mapping children vs. slow-mapping adults: Assumptions about words and concepts in two literatures. *Behavioral and Brain Sciences, 24*, 1112.
Murphy, G. L. (2007). Parsimony and the psychological representation of polysemous words. In M. Rakova, G. Petho & C. Rákosi (Eds.), *The cognitive basis of polysemy*. Fankfurt am main, Germany: Peter Lang Verlag.
Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition, 78*, 165-188.
Nunberg, G. (1979). The Non-Uniqueness of Semantic Solutions: Polysemy. *Linguistics and Philosophy, 3*, 145-184.
Nunberg, G. (1995). Transfers of meaning. *Journal of Semantics, 12*, 109-132.
Nunberg, G. (2004). The pragmatics of deferred interpretation. In L. Horn & G. Ward (Eds.), *Handbook of Pragmatics* (pp. 344 - 364). Oxford, England: Blackwell.
Ortony, A. (1979). *Metaphor and thought*. Cambridge, UK: CUP.
Papafragou, A. (1996). On metonymy. *Lingua, 99*, 169-195.
Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition, 86*, 253-282.

Piaget, J. (1926). *The language and thought of the child*. New York: Harcourt Brace & Co.
Pinker, S. (1979). Formal models of language learning. *Cognition, 7*, 217-283.
Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: HUP.
Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure*. Cambridge, MA: MIT Press.
Pinker, S., Lebeaux, D. S., & Frost, L. A. (1987). Productivity and constraints in the acquisition of the passive. *Cognition, 26*, 195-267.
Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.
Pylkkänen, L. (2008). Mismatching meanings in brain and behavior. *Language and Linguistics Compass, 2*, 712-738.
Pylkkänen, L., Llinas, R., & Murphy, G. L. (2006). The representation of polysemy: MEG evidence. *J Cogn Neurosci, 18*, 97-109.
Pylkkänen, L., & McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (pp. 537–577). New York, NY: Elsevier.
Pylkkänen, L., & McElree, B. (2007). An MEG study of silent meaning. *Journal of Cognitive Neuroscience, 19*, 1905-1921.
Selvin, S. (2004). *Statistical analysis of epidemiologic data*. New York, NY: OUP, USA.
Srinivasan, M., & Snedeker, J. (in prep). Judging a book by its cover and its contents: Evidence for a common representational base underlying polysemous meanings in four-year-old children.
Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development*. Cambridge, UK: CUP.
Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*, 209-253.
Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.
van Walraven, C., & Hart, R. G. (2008). Leave'em Alone-Why continuous variables should be analyzed as such. *Neuroepidemiology, 30*, 138-139.
Vosniadou, S. (1987). Children and metaphors. *Child Development, 58*, 870-885.
Wasow, T. (1981). Comments on the paper by Baker. In C. L. Baker & J. J. McCarthy (Eds.), *The logical problem of language acquisition* (pp. 324-329). Cambridge, MA: MIT Press.

Appendix A: Predicability Questions.

Questions marked with a "2" were used in Experiment 2

Object - Abstract Object

Match Licensed

2 Could a book be thin?

Could a video be plastic?

2 Could a DVD be round?

Could a comic be blue and green?

2 Could a CD be shiny?

Match Unlicensed
2 Could a mystery be about a boy?
Could a show be about animals?
2 Could a movie be an hour long?
Could a story be set in space?
2 Could a song be loud?

Mismatch Unlicensed
2 Could a mystery be thin?
Could a show be plastic?
2 Could a movie be round?
Could a story be blue and green?
2 Could a song be shiny?

Mismatch Licensed
2 Could a book be about a boy?
Could a video be about animals?
2 Could a DVD be an hour long?
Could a comic be set in space?
2 Could a CD be loud?

Object - Event
Match Licensed
Could eating a banana be easy?
2 Could painting a dollhouse be fast?
2 Could reading a book be slow?
2 Could drawing a picture be quick?
Could building a house be difficult?

Match Unlicensed
Could a banana be tiny?
2 Could a dollhouse be wide?
2 Could a book be little?
2 Could a picture be large?
Could a house be small?

Additional Match Unlicensed
Could a girl try eating a banana?
Could a girl finish painting a dollhouse?
Could a woman begin reading a book?
Could a boy start drawing a picture?
Could a man finish building a house?

Mismatch Unlicensed
Could eating a banana be tiny?

2 Could painting a dollhouse be wide?
2 Could reading a book be little?
2 Could drawing a picture be large?
Could building a house be small?

Mismatch Licensed
Could a girl try a banana?
2 Could a girl finish a dollhouse?
2 Could a woman begin a book?
2 Could a boy start a picture?
Could a man finish a house?

Container - Content
Match Licensed
2 Could a pot be cracked?
2 Could a cup be plastic?
Could a kettle be smashed?
2 Could a pitcher be broken?
Could a cauldron be metal?

Match Unlicensed
2 Could some soup be stirred?
2 Could some milk be spilled?
Could some water be boiling?
2 Could some juice be poured?
Could some magic potion be bubbling?

Mismatch Unlicensed
2 Could some soup be cracked?
2 Could some milk be plastic?
Could some water be smashed?
2 Could some juice be broken?
Could some magic potion be metal?

Mismatch Licensed
2 Could a pot be stirred?
2 Could a cup be spilled?
Could a kettle be boiling?
2 Could a pitcher be poured?
Could a cauldron be bubbling?

Control
Match
Could a door be brown?
Could a rock be heavy?

Could a door be tall?
Could a rock be large?
Mismatch
Could a rock be angry?
Could a door be happy?
Could a rock be sad?
Could a door run across the room?