Some puzzling findings regarding the acquisition of verbs

Joshua K. Hartshorne
Yujing Huang
Lauren Skorb
Boston College, USA

Abstract: On the whole, children acquire frequent words earlier than less frequent words. However, there are other factors at play, such as an early "noun bias" (relative to input frequency, toddlers learn nouns faster than verbs) and a "content-word bias" (content words are acquired disproportionately to function words). This paper follows up reports of a puzzling phenomenon within verb-learning, where there appears to be a large effect of argument structure class, such that verbs in one class (experiencer-object verbs) were learned substantially earlier than those in another (experiencer-subject verbs) despite being much lower frequency. In addition to the possibility that the aforementioned results are a fluke or due to some confound, prior work has suggested several possible explanations: experiencer-object ("frighten-type") verbs have higher type frequency, encode a causal agent as the sentential subject, and perhaps describe a more salient perspective on the described event. In three experiments, we cast doubt on all three possible explanations. The first experiment replicates and extends the prior findings regarding emotion verbs, ruling out several possible confounds and concerns. The second and third experiments investigate acquisition of chase/flee verbs and give/get verbs, which reveal surprising findings that are not explained by the aforementioned hypotheses. We conclude that these findings indicate a significant hole in our theories of language learning, and that the path forward likely requires a great deal more empirical investigation of the order of acquisition of verbs.

Keywords: language acquisition, perspective pairs, psych verbs, verb-learning

Corresponding author: Joshua K. Hartshorne, Department of Psychology and Neuroscience, Boston College, 140 Commonwealth Ave., McGuinn 522, Cambridge, MA 02467. Email: joshua.hartshorne@bc.edu.

ORCID ID: https://orcid.org/0000-0003-1240-3598

Citation: Hartshorne, Joshua K., Huang, Y., & Skorb, L. (2023). Some puzzling findings regarding the acquisition of verbs. *Language Development Research, 3*(1), 65–104. https://doi.org/10.34842/2023.535

Introduction

Learning a word is more than just a function of having heard it. On the whole, children do acquire frequent words earlier than less frequent words, but that is hardly the end of the story (Goodman, Dale, & Li, 2008; Hansen, 2017). Controlling for frequency, nouns are learned earlier than verbs, which are in turn learned earlier than closed-class words (Gentner, 1982; Goodman et al., 2008; Hansen, 2017). Controlling for word type, highly-imageable words are learned earlier than less-imageable words, perhaps because it is easier to identify the intended referent during conversation (Hansen, 2017; McDonough, Song, Hirsh-Pasek, Golinkoff, & Lannon, 2011). Indeed, raw frequency may be the wrong measure: rather than word knowledge slowly accumulating with each exposure, word-learning may be disproportionately driven by highly-informative learning opportunities, which give rise to "eureka" moments (Medina, Snedeker, Trueswell, & Gleitman, 2011).

One issue particular to learning verbs is learning how they convey who does what to whom. A child who knows no more about *bite* than what sort of action it describes will be at a loss to distinguish between *dog bites man* and *man bites dog*. To really master a verb, she must know its argument structure: which event roles (biter, bite-ee) are realized in what syntactic positions (subject, direct object). The need to learn argument structure might partly explain why verbs are acquired more slowly than nouns.

Recent findings from Hartshorne, Pogue, & Snedeker (2015) suggested that argument structure may also explain why certain verbs are harder to learn than others. In particular, they found that "psych verbs" that realize the experiencer as the direct object (*A frightened/pleased/angered B*; "frighten-type verbs) are acquired substantially earlier than those that realize the experiencer as the subject (*A feared/liked/hated B*; fear-type verbs), despite being much lower frequency. Specifically, while English-speaking children have already acquired a handful of frighten-type verbs by the age of 4, they do not start acquiring fear-type verbs until about a year later. In particular, although four year-olds use words like *like* and *love* in spontaneous speech, they struggle to distinguish who did what to whom, treating *A loves B* as equivalent to *B loves A* (and similarly for other fear-type verbs). Illustrating just how unexpected this finding was, it had actually been observed in prior studies but never taken seriously, being instead either dismissed as the result of confounds or not remarked upon at all (Bowerman, 1990; Braine, Brooks, Cowan, Samuels, & Tamis-LeMonda, 1993; Messenger, Branigan, McLean, & Sorace, 2012; Tinker, Beckwith, & Dougherty, 1988); by addressing those confounds and reporting multiple targeted studies using different methods, Hartshorne, Pogue, and Snedeker (2015) established the finding as something in need of explanation.

Hartshorne and colleagues interpret this finding as evidence for a "**privileged link**" between causality and sentential subjects. Specifically, a long line of research suggests both a cross-linguistic tendency for the agents of caused events to be mapped onto

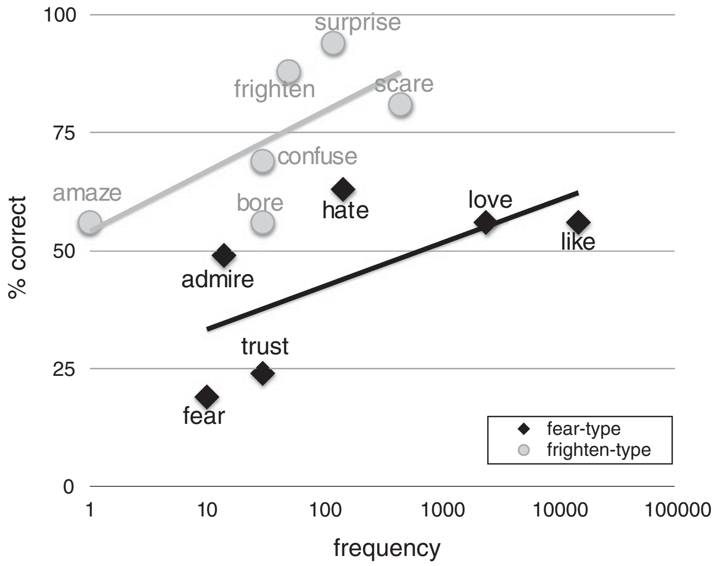

Figure 1. ***Taking frequency in child-directed speech into account, four year-olds were substantially more likely to successfully interpret who did what to whom for frighten-type verbs relative to fear-type verbs (Hartshorne et al., 2015). Figure used with permission.***

subject position and a corresponding learning bias on the part of young children to expect agents of caused events to be subjects of corresponding sentences (Braine, 1992; Dowty, 1991; Fisher, Gertner, Scott, & Yuan, 2010; Levin & Hovav, 2005; Lidz, Gleitman, & Gleitman, 2003; MacWhinney, 1977; Marantz, 1982, 1982; Pinker, 1984; Strickland, Fisher, Keil, & Knobe, 2014). This set of findings has prompted a number of theorists – both Empiricist and Nativist – to argue for an innate bias for agents of caused events to be realized as sentential subjects (or whatever the analog of SUBJECT is in the account adopted by the theorist), and that this bias is a key part of what makes langauge learnable in the first place (for discussion, see Hartshorne et al., 2016). Critically, across a variety of languages, frighten-type verbs do encode causality (*A frightened B* means, roughly, "A caused B to feel fear") whereas fear-type verbs do not (*A feared B* means, roughly, "A was disposed to feel fear about B") (Hartshorne et al., 2016). (Note that while these findings overturned some earlier proposals that treated *fear* and *frighten* as synonymous, differing only in the syntax (Belletti & Rizzi, 1988; Dowty, 1989; Grimshaw, 1990), those had never been tested empirically.) Thus, the "privileged link" would give children a leg up in learning frighten-type verbs, and might even impede the acquisition of fear-type verbs.

The data reviewed above are certainly consistent with the privileged link hypothesis, but you can draw an infinite number of lines through a single point; there are a number of reasonable alternative explanations. One alternative explanation is **type frequency**:

While fear-type verbs have high token-frequency, they are relatively rare as types. English admits around 251 verbs that use frighten-type argument structure but only 49 fear-type verbs (Kipper, Korhonen, Ryant, & Palmer, 2008). The difference becomes even more stark when one considers that frighten-type verbs are a special case of an extremely robust pattern (caused changes of state being realized in transitive syntax with the AGENT as the subject and the PATIENT as the direct object), whereas fear-type verbs are quite unusual (only a handful of other verbs have EXPERIENCERs as subjects). According to a variety of learning theories – including both Nativist and Empiricist – this high type frequency should benefit frighten-type verbs, for instance by helping learners identify the set of features of verbs that reliably predict which verbs admit that particular argument structure (Goldberg, 2006; Pinker, 1989).

A second alternative is **differential salience**: Most utterances are contextually ambiguous: there are many things someone could be talking about (Gleitman, 1990; Quine, 1964). For instance, it is often the case that when there is fearing going on, there is also some frightening; the speaker can choose which to comment on. If children find frightening more salient than fearing, they may be more disposed to successfully identify labeling of frightening than labeling of fearing. Unfortunately, no data currently speak to whether there is a salience asymmetry between frighten-type events and fear-type events.

There are a number of other, less easily testable options. For instance, it is possible that children's exposures to fear-type verbs are relatively short on highly informative "eureka" moments, making them effectively lower-frequency. While there are methods for quantifying the availability of eureka moments, they are time-consuming and difficult to scale (cf. Medina et al., 2011). It may be that the thoughts encoded by fear-type verbs (a habitual disposition towards a particular emotion) are more complex, harder to represent, or emerge later in cognitive development than those encoded by frighten-type verbs (an externally-caused change of emotional state). It is not clear at the moment how this possibility would be tested. It may be that acquisition of frighten-type verbs is aided by the acquisition of other, related argument realization patterns, or that the acquisition of fear-type verbs is impeded by interference from other, contrasting argument realization patterns. This is currently difficult to test because the nature of argument realization patterns remains highly controversial (Levin & Hovav, 2005), and because very little is known about their acquisition outside of a handful of patterns and languages (for comprehensive reviews, see Ambridge et al., 2018, 2020).

There are no doubt other possibilities as well, including of course the possibility that the psych verb findings are a statistical fluke: It happens that children learn their first few frighten-type verbs before learning their first few fear-type verbs, but the reasons are idiosyncratic to each verb and have nothing in particular to do with argument structure class.

Goals of the present study

The present study has two main goals: a) to begin to assess the robustness and generalizability of the finding that argument structure class modulates the timeline of verb acquisition, and b) if the finding is robust and generalizable, tease apart the three hypotheses highlighted above.

A challenge in assessing generalizability is that there are hundreds of argument structure classes in English alone (groups of verbs that use distinct argument structures) (Kipper et al., 2008). Collecting data on even just a substantial portion of them is a long-term project. Unfortunately, we cannot lean much on pre-existing data, since there is almost no prior work identifying the ages of acquisition of the argument structures of specific verbs. Studies of argument structure knowledge typically involve older children who already know quite a few verbs in that class (Ambridge et al., 2018; Pinker, 1989). Studies of vocabulary emergence largely depend on spontaneous production by the child or parental report that the child knows the word, neither of which directly assesses knowledge of argument structure and may in fact be uncorrelated with such knowledge (Hartshorne et al., 2015).

Thus, in order to increase the informativity of the current project, we focused on "perspective pairs". These are groups of verbs which describe similar types of events but contrast in which event-participant is realized as the sentential subject. Psych verbs are an example, where *A feared B* and *B frightened A* can both be said of the same event. While not every fear-type verb has a frighten-type counterpart, the classes as a whole systematically differ in argument structure while describing highly similar types of events.

Focusing on perspective pair classes has the distinct advantage of minimizing uncontrolled differences across verbs. Thus, perhaps children learn *frighten* later than *kick* because the former involves describing a mental state, something that young children struggle with (Wellman, 1992). Or perhaps it is due to any of the myriad other ways that frightening differs from kicking. In contrast, there are far fewer semantic differences between frighten-type and fear-type verbs, and the ones that exist are reasonably well understood (such as how they encode causation).

A second advantage of perspective pairs is that they provide perhaps the best opportunity for testing the salience hypothesis described above. As our science develops, we may eventually have good mechanisms for quantitatively comparing the relative salience of events of frightening vs. events of kicking, but currently this is quite difficult. This question is more straightforward for perspective pairs, and indeed there is some prior work that can inform our investigation (see review below).

We conducted three experiments. The first replicates and extends Hartshorne, Pogue, &

Snedeker (2015) with a larger number of psych verbs (16 vs. 12) and a wider range of ages (3-6 vs. 4-5), while simultaneously allowing us to address some possible concerns about that study's methods. If the prior findings do not generalize at least this much, that would fundamentally change the question. Experiments 2 and 3 focused on chase/flee verbs and give/get verbs, respectively. Chase-type and flee-type verbs differ in terms of whether the pursuer is the subject (*A chased/pursued/followed B*) or an oblique object (*A fled/escaped/ran from B*).[1] Give-type and get-type verbs differ in whether the SOURCE is the subject and the GOAL is an oblique (*A gave/passed/sold B to C*) or *vice versa* (*A got/grabbed/bought B from C*). (Note that, like nearly all verbs, fear/frighten, chase/flee and give/get verbs can appear in other sentence frames as well; here, we focus on the ones that show the contrast most cleanly.)

Critically, these three case studies set up a substantial number of clear comparisons with respect to token frequency, causality, type frequency, and salience (Table 1). The frequency comparisons are straightforward and presented in Tables 2, 4, & 6). We discuss the other three comparisons in detail below.

Table 1: For each of several factors, if that factor was determinative, which verbs would be acquired first.

	Exp1	Exp2	Exp3
pair	fear/frighten	chase/flee	give/get
token frequency	fear	chase	neither
causality	frighten	chase	give
type frequency	frighten	chase	give
salience	??	?chase	get

Causal Semantics

As reviewed above, the "privileged link" for causality hypothesis predicts earlier learning for frighten-type (which have a causal semantics) relative to fear-type (which do not). However, while semantic analysis typically ascribes causality to the subject of both give-type and get-type verbs, and to *neither* chase-type nor flee-type (Kipper et al., 2008; Pinker, 1989), there do not appear to have been any systematic studies (Hartshorne, Bonial, & Palmer, 2014).

[1] While chase-type and flee-type verbs can describe the same events, there is evidence that they – like fear-type and frighten-type verbs – are semantically distinct. Gleitman (1990) reports a personal communication from Steven Pinker arguing that intentional participation in the event is entailed only for the subject of chase/flee verbs: one can chase something that is not fleeing (e.g., a storm) or flee something that is not chasing (e.g., a tsunami). So far as we know, there has not been any systematic study of the verb classes to determine whether this generalizes, though initial inspection suggests that it does. In any case, the exact semantics of these verbs will not be critical for the present study.

We conducted a study in order to obtain quantitative, empirical measurements.[2] The task involved judging who caused various events to happen: specifically, 30 events involving each of the 30 verbs used across the three experiments below (see Tables 2, 4, & 6). Thus, participants were asked to answer questions of the form:

> Who made this happen?: Agnes frightened Beatrice.
>
> Ages, Beatrice, Both of them, Someone else, Nobody (these things just happen), Can't tell

Note that the primary response of interest was how many participants assigned causality to the sentential subject; the other options were included in order to provide natural alternatives.[3]

We recruited native English-speaking adults through Prolific, aiming for 100 participants after exclusions. The final sample was 101 (mean age = 39, range = 18 - 68). An additional 53 participants were excluded for missing one or more catch trials where the answer to the question was stated explicitly (ex: "Who made this happen?: Agnes made Beatrice do something."). There were five catch trials targeting the five primary judgments of interest (all except "can't tell"). This ensured that participants who were included understood roughly what judgments we wished them to make.

As expected, participants were far more likely to judge the subject of frighten-type verbs to have caused the event (88%) than subjects of fear-type verbs (44%) (Table 6). In contrast to prior linguistic analyses, participants also judged the subject of chase-type verbs to be more likely to cause the event (92%) than subjects of flee-type verbs (63%) (Table 4), and the subject of give-type verbs (91%) more than the subjects of get-type verbs (46%) (Table 6). For each of the perspective pairs, the differences between classes were categorical, with the exception of unusually large variability across the four get-type verbs.

Thus if the "privileged link" explains the early learning of frighten-type verbs, we should also expect earlier learning of chase-type vs. flee-type verbs and possibly give-type vs. get-type verbs.

Type Frequency

We assessed type frequency using VerbNet, the largest compendium of verb classes in English (Kipper et al., 2008). In terms of type-frequency, as reviewed above, frighten-

[2] We thank an anonymous reviewer for this suggestion

[3] We did inspect the rates at which participants chose the other answers in order to confirm that the results were sensible, but we did not systematically analyze them, other than the analysis described in footnote 13.

type verbs (N = 251) are far more numerous than fear-type (N = 49).

Chase-type verbs (N = 22) are more common than flee-type, of which there appear to only be around 4. Counting is complicated in that while VerbNet records three groups of chase-type verbs (classes 35.1, 35.3, and 51.6), it does not index flee-type verbs, perhaps because there are so few. *Flee from*, *escape from*, and *retreat from* are included in class 51.1, and *run from* in class 51.3.2, but only in the sense of escaping from a *place*, not an entity. Because flee-type verbs are not indexed in VerbNet, it is possible there are more that we have overlooked, though probably not enough to place the class on par with chase-type. Interestingly, there is a class of verbs (class 52; N = 11) in which the fleer is the subject and the pursuer is the direct object (*A dodged/eluded/evaded B*). Because they take direct objects, they belong in a different syntactic category from chase/flee verbs. We did not investigate them in addition to or instead of chase/flee verbs because they are all low-frequency and unlikely to be encountered by children.

Finally, give-type verbs (N = 82; classes 13.1, 13.2, 13.4.1, 13.4.2) are more numerous than get-type verbs (N = 57; classes 13.5.1, 13.5.2, 57), but the imbalance is not as stark as for the other perspective pair classes.[4]

Salience

As already noted, there is no evidence bearing on the question of whether frighten-type event construals are more or less salient than fear-type construals, and we do not investigate it here (it is currently clear how to do so). With regards to chase/flee, it has been argued that children and adults are biased to encode ambiguous events in terms of chasing rather than fleeing (e.g., Landau & Gleitman, 2015). However, the evidence is thin and mixed. Fisher, Hall, Rakowitz, and Gleitman (1994) indeed found that three year-olds and adults described one ambiguous chase/flee scene in terms of chasing rather than fleeing. However, Gleitman, January, Nappa, and Trueswell (2007) ran a similar study with adults using two stimuli, finding a chase-bias for one and a flee-bias for the other. There do not appear to be any other empirical studies.

There is a more robust literature indicating that when presented with an ambiguous give/get event, both children and adults focus on the getting over the giving. Children and adults are more likely to remark on and remember GOALS than SOURCES (Freeman, Sinha, & Stedmon, 1981; Fujita, 2000; Lakusta & Landau, 2005, 2012; Papafragou, 2010; Regier & Zheng, 2007). This suggests that get-type verbs should be privileged for two reasons. First, the GOAL is obligatory for get-type verbs and optional for many give-type verbs (*A sent/sold/passed B [to C]*). Second, it is widely argued that more salient entities

[4]We exclude verbs of future having (*A awarded/bequeathed/owed B to C*) from give-type verbs. The semantics are critically different in that there is no caused change of possession, just a promised change of possession. If these are included in the count, the advantage in token-frequency for give-type verbs becomes more imbalanced at 110 to 57.

are preferentially mapped onto sentential subjects.

In terms of language acquisition, there is only limited evidence as to an advantage for get-type (or give-type) verbs. In studies looking at spontaneous description, participants often avoid both give-type and get-type verbs, instead preferring THEME-subject verbs (*A walked/ran/rolled to B*). Lakusta and Landau (2005) similarly found that children were more likely to use known give-type verbs than get-type verbs when describing ambiguous scenes (Lakusta & Landau, 2005).[5] A smaller novel-word learning study showed a similar bias towards give-type verbs (Fisher et al., 1994). However, Bowerman (1990) reports in a diary study of two children that the first get-type verb (*get*) emerges in spontaneous speech only just prior (1;8) to the first give-type verbs (*give/gimme*, *tell*, and *read me*, all of which emerge at 1;9). Since get-type verbs appear first but give-type appear in larger numbers, this seems to be a draw.

Summary of Predictions

The predictions for the three experiments are summarized in Table 1.

- If token frequency is the key factor, we would expect earlier acquisition of fear-type verbs compared to frighten-type; chase-type compared to flee-type; and roughly equal learning for give-type and get-type. This outcome seems unlikely given the results of Hartshorne, Pogue, and Snedeker (2015), but perhaps our replication-and-extension will show different results.
- If the "privileged link"/causality hypothesis is correct, we should see earlier acquisition of frighten- type than fear-type verbs, chase-type verbs than flee-type, and give-type relative to get-type.
- If type frequency is the key factor, we should see earlier acquisition of frighten-type than fear-type verbs; chase-type relative to flee-type; and give-type relative to get-type.
- If the salience hypothesis is correct, we have no strong predictions for psych verbs, but have a weak prediction of earlier learning of chase-type than flee-type verbs, and a stronger prediction of earlier learning of get-type than give-type.

By using these three case studies, we hope to disentangle the four main hypotheses under consideration.

[5]Lakusta and Landau (2005) argue that the salience hypothesis predicts a preference for give-type verbs. In particular, they posit that children should prioritize the mapping to the prepositional phrase and thus find it easier to map the more salient GOAL onto the prepositional phrase, rather than the SOURCE. This is an intriguing notion, but it is incompatible with most prominent theories of language acquisition, and it remains to be seen whether a new theory can be constructed around it. It is also inconsistent with the data we present below. Thus, we do not consider this hypothesis further in this paper.

Overview of analyses

In all three experiments, we submit the data to a mixed effects logistic regression with centered main effects of verb type, log frequency, and subject age in months.[6] In order to improve convergence and avoid issues with singularity, we fit the model with partially-Bayesian regression with Wishart priors on the covariance matrix for random effects, using the blme package with bobyqa optimization (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013).

We included random intercepts by subject and verb as well as a random slope of verb type by subject. We chose not to use a maximal random effects structure for three reasons. First, there is some debate about whether doing so is even desirable (Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). In particular, specifying overly complex models may substantially increase Type II error. Indeed, we found that for several of our analyses, a maximal random effects structure resulted in *no* significant effects, not even effects of age or log frequency, even though these are extremely well attested in the literature and clearly visible in our data. Second, the small number of verbs (ranging from 6 to 16 per study) means that random slopes are estimated based on very few data points. For instance, estimating a random slope for participants for the interaction of verb type and log frequency involves calculations of slopes based on as few as 3 and at most 8 boolean responses. Third, empirically we found that in most cases random slopes led to worse fits (as assessed by BIC and AIC) – when fitting could even be done successfully. Analysis of the random effects structure for Exp. 3 did justify random slopes of verb type and log frequency by subject, but we felt using a different random effects structure for different experiments would impede comparison across experiments. Thus, we settled on a relatively simple random effects structure with only one random slope (verb type by subject), which is a fairly simple slope based on a relatively large amount of data.

While this is our primary analysis, we considered two other analysis methods. First, we fit the primary models using Bayesian regression as implemented in brms (Bürkner, 2017, 2018) and calculated Bayesian p-values with the p_map function from bayestestR (Makowski, Ben-Shachar, & Lüdecke, 2019).[7] Unless data sets are very large, Bayesian analyses tend to be biased in favor of the null hypothesis, so it is not surprising that some of the effects that were significant in our main analyses are not significant in the Bayesian regressions, though for the most part the key results remained the same. The results of these models are reported in footnotes (in general, we use footnotes in

[6]Not centering these variables frequently led to expected failures to converge.

[7]For the intercept, fixed effects, and random effects, we used the relatively informative prior of a half-normal distribution centered at 0 and with a standard deviation of 1. We ran six chains of 8000 samples each, including 1000 warmup samples. For two of the models, we raised alpha_adapt to 0.9 in order to address a small number of divergent transitions.

lieu of supplementary materials). Second, we used model comparison to prune fixed effects from the primary model. Unfortunately, we obtained wildly different results using AIC, which generally favored retaining most or all of the fixed effects, and BIC, which is a very conservative method – particularly when data sets are not very large – and which generally favored eliminating most or even all of the fixed effects. Because the effects of log frequency and age are well-established and quite clear in the data, this suggests BIC is overly conservative. We do not describe these results in detail, but they are memorialized in our reproducible RMarkdown document, available at osf.io/k5xud. Because the Bayesian regression and BIC method both appear to be overly conservative, we do not consider their results as strongly indicative of the null, but it certainly does mean that the evidence we present below is not incontrovertibly strong.

We also report follow-up analyses for each experiment, focused on each individual age group (3 year-olds, 4 year-olds, 5 year-olds, and 6 year-olds).

R packages used to prepare this reproducible document include *papaja* (Aust & Barth, 2020), *knitr* (Xie, 2015), *ggplot2* (Wickham, 2016), *ggeffects* (Lüdecke, 2018), and *stargazer* (Hlavac, 2018).

Experiment 1: Fear/Frighten (Psych) Verbs

This experiment replicated the method of Exp. 1 of Hartshorne, Pogue, and Snedeker (2015), with the following changes: a) we added two new fear-type verbs and two new frighten-type verbs, b) we expanded the age range to 3 to 6, c) in order to accommodate the attention spans of the younger participants, we split the verb lists so that each subject saw only half of the stimuli, and d) we presented half the participants with critical verbs in the past tense, half in present tense.

The manipulation of tense requires some explanation. Tense affects fear-type and frighten-type verbs differently. While fear-type verbs refer to a habitual state whether used the present tense (*A fears B* [*always / *just the once*]) or past tense (*A feared B* [*always / * just the once*]), frighten-type verbs most naturally refer to a habitual or repeated event in the present tense (*A frightens B* [*always / *just the once*]) and a single event in the past tense (*A frightened B* [*?always / just the once*]). Unfortunately, Hartshorne, Pogue, and Snedeker (2015) did not explicitly control the tense used, so it is unclear whether the same tense was used for all verbs. It is unlikely this could explain poorer performance on fear-type verbs, which are unaffected by tense, but out of an abundance of caution, Exp. 1 included a tense manipulation in order to test the question directly.

Method

All research reported in this paper was approved by the Boston College and Harvard University Institutional Review Boards.

Participants

We recruited 290 native English-speaking children between the age of 3 and 6 years from the Greater Boston Area from parks, museums, and preschools. Of these, we excluded 47 children due to coding error or failure to complete more than 50% of the experiment or both. We had intended to have 64 participants per age group (one per list; see below), but delays due to the pandemic made it impossible to reach the full complement in some cases. In others, we ended up with more participants than intended for the reasons described in "Procedure". The final numbers included 54 3 year-olds, 73 4 year-olds, 82 5 year-olds, and 34 6 year-olds.

Materials

A total of 16 verbs were tested (8 fear-type and 8 frighten-type), including the 12 verbs from Hartshorne, Pogue, and Snedeker (2015) and 4 additional verbs. Part of the goal was to include three relatively high-frequency verbs (*enjoy*, *dislike*, *bug*) that were overlooked in construction of the earlier experiments (Hartshorne, Pogue, and Snedeker (2015) did not have access to as complete a list of fear/frighten verbs as we do). The final verb (*anger*) is relatively low-frequency but was commonly used by children during a pilot verb-elicitation study and so was included.

We estimated frequencies for fear/frighten verbs in speech directed to children ages 36 to 84 months (the age range in the present studies) by using childesr (Sanchez et al., 2019) to aggregate all speech (other than that by the target child) in North American English corpora in CHILDES (MacWhinney, 2000) where the target child was in the age range and where part-of-speech tags were available. This aggregate corpus consisted of 3,044,358 tokens. Frequencies for our stimuli are shown in Table 2. A disproportionate number of high-frequency words were fear-type, though the difference between types across the 16 verbs did not reach significance ($\Delta M = 0.19$, 95% CI $[-2.14, 2.52]$, $t(11.44) = 0.18$, $p = .863$).[8]

Table 2: Verbs used in Experiment 1, with frequency in parts per million and probability that causality is assigned to the sentential subject.

Verb	Type	Frequency	LogFrequency	SubjCause
dislike	fear	0.3	0.3	50
admire	fear	1.6	1.0	48
fear	fear	2.3	1.2	27
trust	fear	4.6	1.7	45
enjoy	fear	48.9	3.9	40
hate	fear	55.2	4.0	48

[8]For these and other statistics, we used log-transformed frequencies.

Verb	Type	Frequency	LogFrequency	SubjCause
love	fear	269.0	5.6	47
like	fear	2246.5	7.7	48
anger	frighten	0.0	0.0	86
bug	frighten	8.9	2.3	87
frighten	frighten	15.8	2.8	90
confuse	frighten	16.4	2.9	83
amaze	frighten	20.0	3.0	94
bore	frighten	27.9	3.4	78
surprise	frighten	56.5	4.1	95
scare	frighten	214.2	5.4	94

For each of the 16 verbs, we created four scenarios, counterbalancing the pair of characters involved and which member of the pair experienced the emotion. We created two stimulus orders as follows: We did a mid-line split on token frequency for both fear-type and frighten-type verbs and placed the 4 highest-frequency of each type in the first half the list. We then created a second list reversing the orders *within* each half, so that the high-frequency verbs remained in the first half. The purpose of this was to allow participants to continue on to the second half if they have sufficient interest, but this rarely happened. Ultimately, we began randomly assigning participants to one half or the other. (For those participants from the initial phase of testing who had completed both halves, we excluded the second half.) The result was 64 lists: 2 (item set) by 2 (present vs past) by 2 (animal pair) by 2 (animal roles) by 2 (item order) by 2 (target response) design.

During testing, we discovered an error in the lists such that for four of the low-frequency verbs (*bug, anger, enjoy, dislike*) only one of the character pairs was used. We fixed the lists and began replacing those participants (N=27) who had not seen the intended character pairs. However, because we were unable to finish replacing them due to the pandemic, we included both the original and "replacement" participants in order to improve power. Nonetheless, excluding the to-be-replaced participants has no effect on the qualitative pattern of results.

Procedure

Participants were run one at a time. An experimenter read stories accompanied by pictures. After each story, a second experimenter used a Mickey Mouse puppet to say what Mickey thought happened in the story. Sometimes what Mickey said was correct and sometimes it was incorrect. If the participant thought that what Mickey said was correct, they gave Mickey a cookie. If they thought that what Mickey said was incorrect, they gave him coal. The experimenter began with two practice stories involving action

Table 3: Regression results for psych verbs

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.353	0.074	4.781	0.00000
Age	0.290	0.056	5.166	0.00000
VerbType	0.248	0.144	1.718	0.086
LogFrequency	0.246	0.082	3.020	0.003
Age:VerbType	0.380	0.108	3.530	0.0004
Age:LogFrequency	0.113	0.060	1.885	0.059
VerbType:LogFrequency	0.346	0.162	2.138	0.033
Age:VerbType:LogFrequency	0.196	0.119	1.649	0.099

verbs (*hug* and *kiss*), followed by the eight test trials. Participants received explicit feedback on the two practice trials but only general affirmative reactions to the critical trials. Responses were coded on site by a second experimenter. When possible, these were then double-checked from a video recording.

Results

Data for this and all experiments are available at osf.io/k5xud.

The dependent measure was accuracy: correctly accepting a true statement or rejecting a false one. We conducted a preliminary analysis to assess whether tense systematically affected the results. We submitted the results to a partially-Bayesian mixed effects logistic regression with Wishart priors on the covariance matrix for random effects, using the blme package with bobyqa optimization (Chung et al., 2013).[9] We included main effects of verb type and tense, along with their interaction, and random slopes of verb type by subject and tense by verb.[10] The main effect of tense was not significant (B = -0.13, SE = 0.15, Wald's z = -0.88, *p*=0.38), nor was the interaction of tense and type (B = 0.17, SE = 0.29, Wald's z = 0.59, *p*=0.56). Thus, all subsequent analyses collapsed across tense.

[9] In the interests of full communication of statistics, we provide more information rather than less. This includes providing exact p-values for all significant results, in order to support meta-analysis. For reasons of space, non-significant p-values are not reported except for marginal values ($.05<p<0.10$). While there is some debate over how to interpret marginal effects, we have erred on the side of providing information that readers may find useful. In any case, none of the marginal effects reported directly impinge on our conclusions one way or another.

[10] This is actually a maximal design. The random slope of verb type by subject is included in all analyses, so its inclusion requires no additional justification. The random slope of tense by verb is a fairly simple slope (the effect is categorical) and Exp. 1 has a relatively large number of verbs (16). In any case, excluding this random slope does not change the pattern of results. Note that the random slope of tense is not included in other analyses because it is not relevant to those analyses.

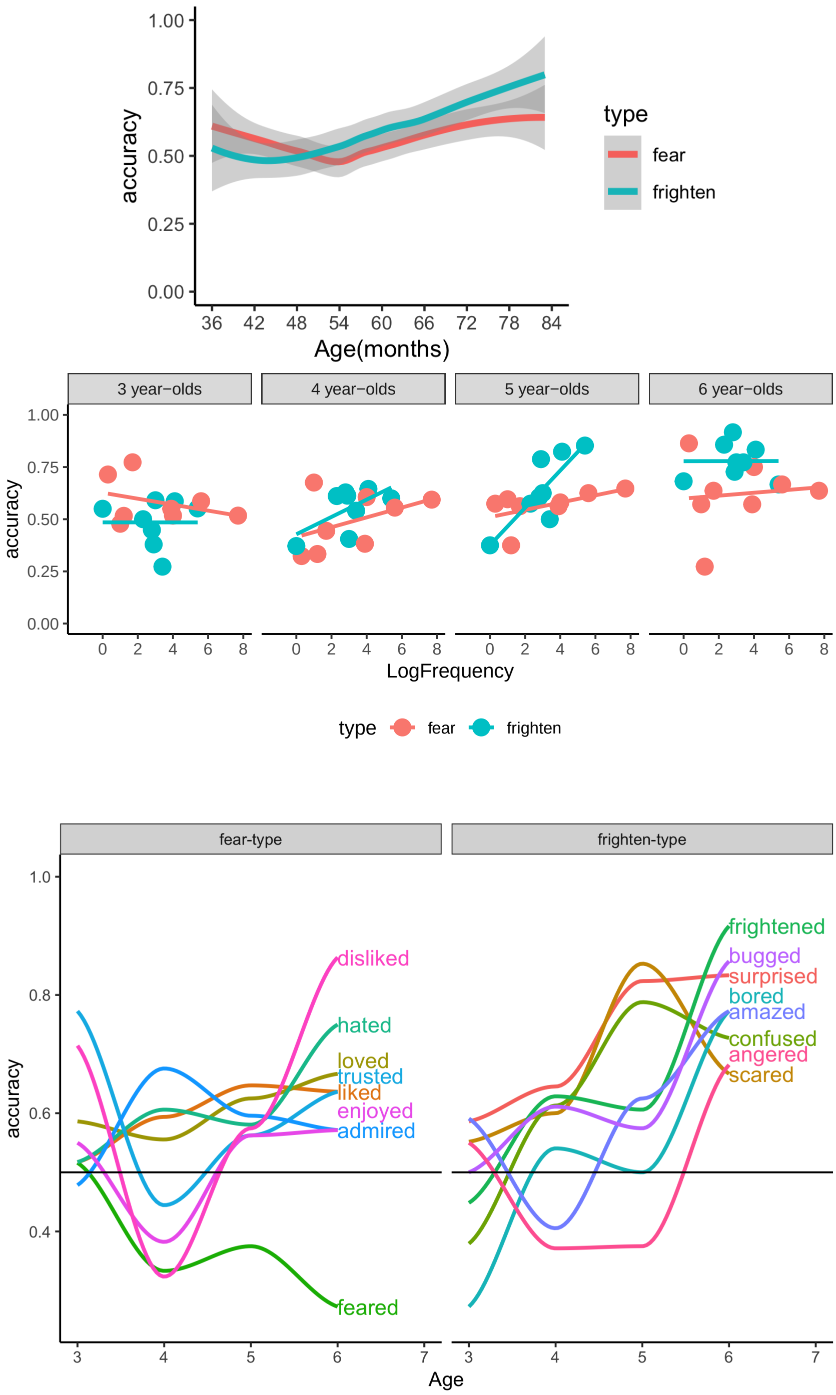

Figure 2. ***Results from Exp. 1. Top: Averaging across verbs within type, with LOESS smoothing over age. Middle: Accuracy for each verb against log frequency, split into four age groups, with linear regressions shown. Bottom: Performance on each verb across ages, aggregated into four age groups, with LOESS smoothing.***

We then conducted our main analyses, as described above in "Overview of Analyses." Results are shown in Table 3. The interaction of verb type by log frequency was significant, reflecting a larger effect of frequency on frighten-type verbs, as was the interaction of verb type by age, reflecting a larger advantage of frighten-type verbs in older participants.[11]

As explained in "Overview of Analyses," we followed up this analysis by considering each age group separately. For 3 year-olds, neither the main effects nor their interaction were significant (ps ≥ 0.12). For 4 year-olds, the only effect to approach significance was that of log frequency (B = 0.27, SE = 0.14, Wald's z = 1.90, p=0.057). By 5, there was a significant effect of log frequency (B = 0.49, SE = 0.13, Wald's z = 3.80, p=0.00014), a marginal effect of verb type (B = 0.41, SE = 0.23, Wald's z = 1.80, p=0.073), and a significant interaction (B = 0.71, SE = 0.25, Wald's z = 2.80, p=0.0055). By the age of 6, the only significant effect was that of verb type (B = 1.20, SE = 0.55, Wald's z = 2.10, p=0.033). Note that the sample sizes for 3 year-olds and 6 year-olds were smaller, perhaps explaining the fewer number of significant effects.

Fig. 2 visualizes these results along three dimensions. There is a clear overall advantage for frighten-type verbs beginning at around 50 months (Fig. 2 top). Interestingly, while there is a clear upward trajectory for frighten-type verbs beginning at around 48 months, there is no clear pattern for fear-type verbs, on which performance at 84 months is similar to that at 36 months.

Fig. 2 (middle) provides another window into the pattern of results: children exhibit higher performance on frighten-type verbs, controlling for frequency, by 4 years old, with the effect growing substantially by 6 years old. This figure is consistent with our age-group-specific analyses: 3 and 4 year-olds appear to be close to chance on most items, though there is a hint of better performance for high-frequency verbs by the age of 4. 5 year-olds show a clear pattern of success largely restricted to high-frequency frighten-type verbs. By the age of 6, children are doing well on all the frighten-type verbs but performance remains low for most fear-type verbs.

The apparent lack of improvement with age on fear-type verbs may be due to learners misclassifying low-frequency fear-type verbs as frighten-type (Fig. 2 Bottom). The highest-frequency fear-type verbs (*like, love, hate*) may not have reached the performance levels of frighten-type verbs, but they did show gradual improvement across the age range. In contrast, four of the lower-frequency verbs (*enjoy, trust, dislike, fear*) actually showed a *decline* from 3 to 4 years old, with all four verbs actually declining to below-chance levels. This matches the results of Hartshorne, Pogue, and Snedeker (2015), who reported significantly below-chance performance on *trust* and *fear* (*enjoy*

[11]The Bayesian regression revealed significant effects of age (p <0.00000001) and log frequency (p=0.034), and a significant interaction of age and verb type (p = 0.0023). The only result significant in the main analyses but not the Bayesian analyses was the interaction of frequency and verb type (p = 0.16)

and *dislike* were not tested).

Discussion

The results of Exp. 1 confirmed the results of Hartshorne, Pogue, and Snedeker (2015) with more verbs and across a larger age range: children began learning frighten-type verbs by the age of 4-5, whereas fear-type verbs remained largely at near-chance levels even at the age of 6. As suggested by Hartshorne, Pogue, and Snedeker (2015), this appears to reflect a tendency to misanalyze fear-type verbs as being frighten-type verbs.

These results are consistent with both the privileged link hypothesis and the type frequency hypothesis. It is unknown how it matches the salience hypothesis, because we do not know whether fear-type or frighten-type event representations are more salient.

Experiment 2: Chase/Flee

Experiment 1 confirmed the prior findings by Hartshorne and colleagues (2015), in which frighten-type verbs were learned earlier than fear-type verbs, despite being lower-frequency. As reviewed above, these admit several difference explanations: frighten-type verbs encode a causal agent as the subject, are more numerous, and arguably encode more salient events than do fear-type verbs.

In order to start disentangling these possibilities, we next considered a different perspective pair: verbs that describe chasing and verbs that describe fleeing. Predictions are summarized in Table 1. As with psych verbs, the class that is the most numerous (chase-type) is also the one where the verb's subject most clearly encodes causality. Thus, if either of these factors explained the early learning of frighten-type verbs, we would expect chase-type verbs to be similarly advantaged. In addition, the chase-type verbs might be early-learned because they arguably encode the more salient event perspective.

One concern about Experiment 1 is that the Truth Value Judgment task may have been more difficult for the youngest children. In particular, describing internal emotional states of the characters requires a fairly involved story (at least, by the standards of stories for 3 year-olds). In contrast, the chase/flee verbs lend themselves naturally to an act-out task – something that was obviously not possible for psych verbs. Thus for Experiment 2, we adopted an act-out task paradigm.

Method

Participants

We recruited 208 children ages 3 through 6 from the Greater Boston Area. All participants were native English speakers. We excluded 41 children for failing to complete more than half of the items or experimentor error. Although we aimed for 40 participants per age (10 per list; see below), we ultimately obtained 43 3 year-olds, 43 4 year-olds, 40 5 year-olds, and 41 6 year-olds.

Materials

Table 4: Verbs used in Experiment 2, with frequency in parts per million and probability that causality is assigned to the sentential subject.

Verb	Type	Frequency	LogFrequency	SubjCause
pursue	chase	0.0	0.0	96
chase	chase	75.9	4.3	92
follow	chase	90.3	4.5	88
flee from	flee	0.3	0.3	57
escape from	flee	4.6	1.7	70
run from	flee	8.5	2.3	61

We selected the 3 highest-frequency chase-type verbs (*chase*, *follow*, *pursue*) and 3 highest-frequency flee-type verbs (*flee*, *escape*, *run*) (Table 4). Frequencies were determined as in Exp. 1, with the caveat that we restricted frequency counts for *run* to specifically the bigram *run from*. Two puppet participants, Giraffe and Tiger, were used for each sentence. We created 6 stories which participants needed to act out with the two puppets. Each story was one sentence long (e.g., *Giraffe chased Tiger*) to minimize demands on working memory and clearly isolate the verb of interest. Giraffe and Tiger were counterbalanced between acting as the subject and the object for each verb. Four lists were created by counter-balancing across lists whether Giraffe or Tiger was the subject of each verb and by creating two item orders, one of which was the reverse of the other.

Procedure

The researchers began by familiarizing the participants with the puppets. Participants were then informed that they would be listening to stories about Giraffe and Tiger, and that they should act out those stories using the puppets. The stories were read aloud, and the researchers recorded whether or not the child correctly demonstrated the meaning of the verb using the puppets. This was operationalized by the child moving

Table 5: Regression results for chase/flee verbs

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.356	0.277	4.897	0.00000
Age	1.002	0.185	5.416	0.00000
VerbType	-1.706	0.502	-3.400	0.001
LogFrequency	1.661	0.316	5.258	0.00000
Age:VerbType	-0.659	0.284	-2.318	0.020
Age:LogFrequency	0.907	0.165	5.502	0.00000
VerbType:LogFrequency	-1.079	0.635	-1.699	0.089
Age:VerbType:LogFrequency	-1.025	0.322	-3.180	0.001

the puppets in the correct direction (i.e. the Giraffe towards the Tiger for “Giraffe chased Tiger”) or by reorienting the chaser to be facing the correct direction (i.e. the Giraffe faced the Tiger for “Giraffe chased Tiger”). The on-site record was later checked by another researcher using a video recording of the study.

Results

The dependent measure was accuracy: whether the participant correctly acted out the event (see description of coding procedure above). The mixed effects logistic regression revealed that every main effect and interaction was significant except the interaction of verb type and log frequency, which trended towards significance (Table 5).[12]

Figure 3 plots the main results. While accuracy was roughly similar on both argument structure classes across the age range studied (Fig. 3, top), this belies a large difference in frequency. In particular, although *chase* and *follow* are much higher frequency than *escape from* and *run from*, they are learned no earlier (Fig. 3, middle). As noted above, this not-quite-significant interaction is actually modulated by a significant three-way interaction of age, token frequency, and argument structure class. To aid interpretation of the three-way interaction between verb type, age, and log frequency (Table 5), we plotted marginal effects (Fig. 4). This shows that at a range of frequencies, the inferred rate of learning for flee-type verbs exceeds that of chase-type. Indeed, the inferred rate of learning for *run from* (Log Frequency = 2.3) is actually higher than that of *chase* (Log Frequency = 4.3) or *follow* (Log Frequency = 4.5).

[12]The results of the Bayesian regression were reasonably similar. In particular, the critical three-way interaction of age, log frequency, and verb type was again significant (p = 0.016). The biggest difference was that the main effect of verb type was not significant (p 0.34), nor was the interaction of age and verb type (p 0.13). However, this was less of a difference in the point estimates than just a high degree of uncertainty. The other effects that were significant in the frequentist analyses where either significant in the Bayesian analyses – the two-way interaction of age and log frequency (p <0.00000001) and the main effect of age (p <0.00000001) – or trended towards significance – log frequency (p = 0.074).

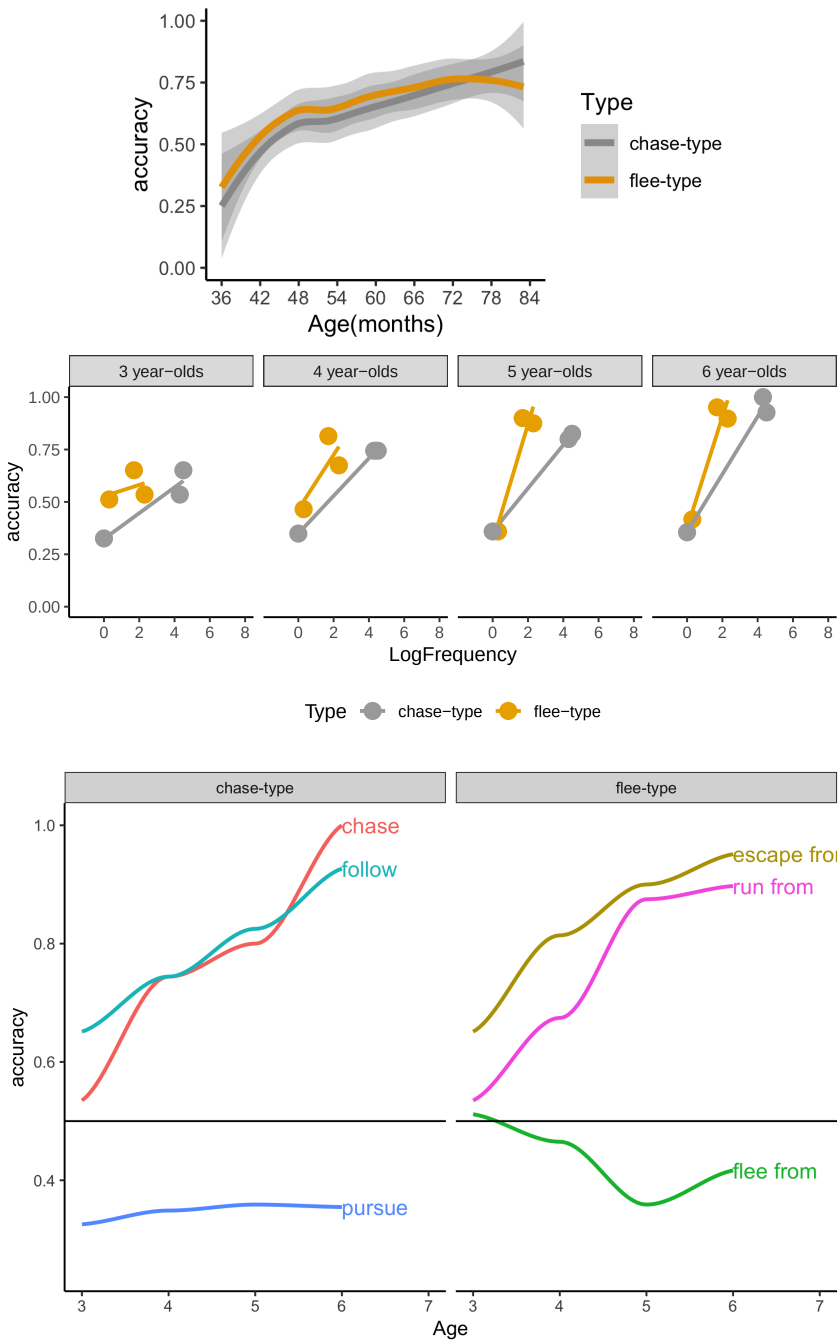

Figure 3. ***Results from Exp. 2. Top: Averaging across verbs within type, with LOESS smoothing over age. Middle: Accuracy for each verb against log frequency, split into four age groups, with linear regressions shown. Bottom: Performance on each verb across ages, aggregated into four age groups, with LOESS smoothing.***

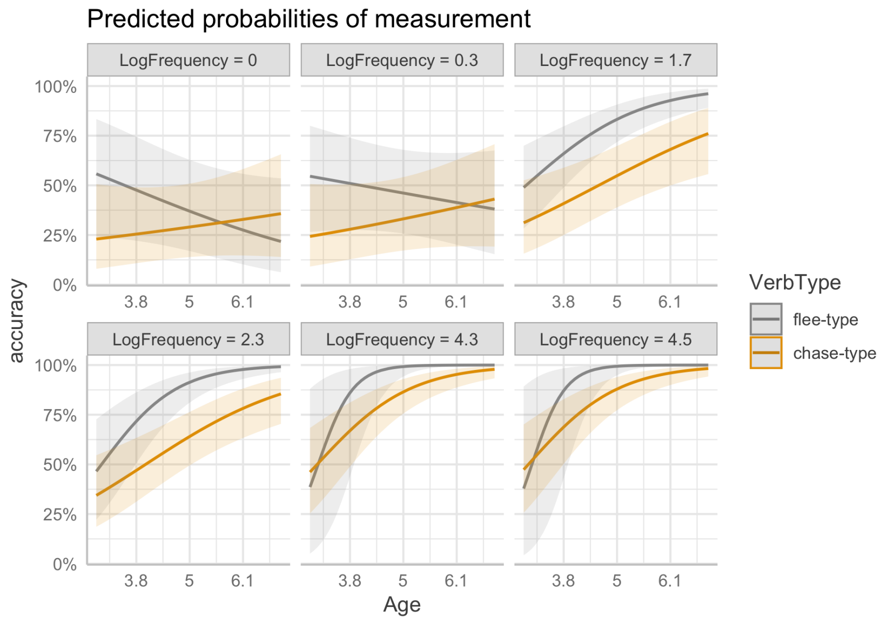

Figure 4. ***Marginal effects for the interaction of verb type, (z-scored) age, and frequency in chase/flee verbs (Exp. 2). Note that the curves plotted are inferred by the model; they do not represent direct observations. For instance, there was no flee-type verb with a log frequency of 0.***

As explained in "Overview of Analyses," we followed up this analysis by considering each age group separately. For 3 year-olds, neither the main effects nor their interaction were significant (*ps* $\geq$ 0.11). For 4 year-olds, there were significant main effects of log frequency (B = 1.30, SE = 0.43, Wald's z = 3, *p*=0.0025) and verb type (B = -1.40, SE = 0.69, Wald's z = -2, *p*=0.045), though their interaction was not significant (B = -0.50, SE = 0.86, Wald's z = -0.59, *p*=0.56). At age 5, there remained significant main effects of log frequency (B = 2.90, SE = 0.64, Wald's z = 4.50, *p*=0.0000075) and verb type (B = -2.70, SE = 1.10, Wald's z = -2.40, *p*=0.015), and the interaction approached significance (B = -2.30, SE = 1.30, Wald's z = -1.80, *p*=0.076). By the age of 6, all effects were significant: log frequency (B = 15, SE = 3.50, Wald's z = 4.20, *p*=0.000029), verb type (B = -13, SE = 4.20, Wald's z = -3, *p*=0.0029), and their interaction (B = -24, SE = 5.80, Wald's z = -4.10, *p*=0.000042).

Discussion

The results of Exp. 2 indicate an early advantage of flee-type verbs relative to chase-type verbs, holding frequency constant, albeit perhaps not as pronounced as that for frighten-type verbs relative to fear-type verbs. All three of our main hypotheses predicted an advantage for chase-type verbs, though in the case of the salience hypothesis, the evidence driving this prediction is weak. In any case, the results were exactly the opposite.

Interestingly, one chase-type verb and one flee-type verb each exhibited stubbornly sub-chance levels: Children responded as if *pursue* was flee-type and *flee from* was chase-type – something that had not resolved even by the age of 6. This is puzzling, and we do not have much to say about it at the moment other than that it is not obviously explicable under any of the hypotheses being considered.

It should be noted that the observed differences between chase-type and flee-type verbs are due primarily to four verbs (the differences in learning outcomes for *pursue* and *flee from* are fairly similar). Thus, different results for a single verb would have substantially changed the statistical outcomes. Unfortunately, because flee-type verbs are so rare type-wise, not much can be done about this other than to consider other perspective pairs as well.

Experiment 3

The results of Experiment 2 converged with those of Experiment 1 in that token frequency was a poor predictor of learning. Otherwise, the results contrast: unlike Experiment 1, the results of Experiment 2 were inconsistent with the suggestion that acquisition is heavily influenced by class token frequency or causality. The results moreover suggested that salience plays little role.

Nonetheless, Experiment 2 considered a very small number of verbs. Thus, in Experiment 3, we turn to another case study: give/get verbs. In addition to being more numerous, give/get verbs offer the advantage of disentangling the predictions of the causal semantics and token frequency hypotheses on the one hand from those of the salience hypothesis on the other (Table 1).

While give and get events can be acted out, it is a bit more complicated to manage with small hands, as there are three entities to keep track of. Moreover, we wished to avoid providing clues to the direction of motion in the form of prepositions (giving *to* vs. getting *from*). Thus, we used a modified video-description task: children watch an event involving a boy and a girl exchanging items and were queried as to what the one of the characters got/gave/etc.

Method

Participants

We recruited 452 native English-speaking children ages 36 to 83 months old from the Greater Boston Area. We excluded 28 children due to experimentor error or failure to complete more than half the trials. While we had intended to recruit 9 participants per age group per list (see below), we fell 8 participants short due to recruitment restrictions. Thus, we finished with 424: 107 3 year-olds, 105 4 year-olds, 109 5 year-olds, and 104 6 year-olds.

Materials

The four give-type verbs and four get-type verbs were selected from the Verbnet Unified Verb Index. These were the four highest-frequency verbs that we were able to identify for each type (Table 6). The two groups did not differ significantly in terms of frequency ($\Delta M = 0.17$, 95% CI $[-4.32, 4.67]$, $t(4.34) = 0.10$, $p = .921$).

Table 6: Verbs used in Experiment 3, with frequency in parts per million and probability that causality is assigned to the sentential subject.

Verb	Type	Frequency	LogFrequency	SubjCause
receive	get	4.6	1.7	4
grab	get	65.4	4.2	96
buy	get	423.4	6.1	58
get	get	6944.0	8.8	26
sell	give	62.1	4.1	73
send	give	75.2	4.3	99
pass	give	85.7	4.5	94
give	give	1318.2	7.2	96

We constructed videos depicting one male and one female experimenter exchanging objects with one another. Fig. 5 shows an example: the man gives an apple to the woman, who then reciprocates by giving a hammer to the man. The participant would then be asked either *what did the boy give?* or *what did the girl give?* The fact that we could query either character allowed us to counter-balance SOURCEs and GOALs within the same stimuli. We counter-balanced another way as well: we made pairs of videos that differed in which character moved first, which should wash out any bias towards the first- (or last-) mover. Similarly, the videos were designed to allow querying one give-type verb and one get-type verb: the same video was used for either *give* or *get*, another for *send* or *receive*, and yet another for *buy* and *sell*. An exception to this design was forced by *grab*, which was paired with *pass*. For obvious reasons, the same video could not be used for both. Moreover, we found videos with reciprocal grabbing to be confusing. Thus, rather than have the man and woman both grab from each other within the same video, we created two videos – one with grabbing by the man and one with grabbing by the woman – and placed them one above the other. Instead of counter-balancing which character acted first, we counter-balanced which video was on top. Because *pass* was matched with *grab* for purposes of counter-balancing, we constructed the *pass* stimuli in a similar way.

Four lists were constructed by counterbalancing the order in which the verbs were

Figure 5. ***Stills from one of the videos for* give*, used in Exp. 3.***

queried and, for a given verb, which character was queried and which character moved first in the video. To achieve this, we used a single fixed order of the videos (which a caveat describes below), counter-balancing which character is asked about and which verb in each verb pair was queried. The caveat is that this was obviously not possible for *pass* and *grab*. Instead, for these we swapped the order of the videos (so on two lists, *pass* was presented before *grab*, and the reverse was true for the other two lists) and the vertical positioning of the videos (see above).

Finally, these four lists were triplicated by making two more sets of videos, each with a different pair of actors and different sets of objects, for a total of 12 lists.

Procedure

Videos were presented on an iPad using Keynote. Prior to watching each video, participants viewed the opening frame and were asked to point to the target objects, with corrections provided as necessary. This helped ensure they could identify the objects well enough to answer the subsequent questions. The video was then played, and the participant was asked the question involving the give/get verb for that video.

Results

The dependent measure was accuracy: whether the participant named the correct item. The three-way interaction of age, log frequency, and verb type was significant, as was the interaction of age and verb type and the main effects of frequency and age (Table 7).[13] The data plots provide context. Averaging within verb-type, there is a clear early advantage for get-type verbs, which disappears by age 6 in part because performance improves with age for give-type but not get-type verbs (Fig. 6 top). Plots of individual verbs (Fig. 6 bottom) show steady improvement with age for all give-type verbs, with

[13]The three-way interaction that was significant in the frequentist analyses only trended towards significance in the Bayesian regression (p=0.08). The interaction of age and verb type was again fully significant (p=<0.00000001), as was the main effect of age (p=<0.00000001). The only effect that was significant in the frequentist analyses that was was clearly not significant in the Bayesian analyses was the main effect of log frequency (p=0.19), though as usual this reflected high uncertainty rather than certainty that there is no effect.

Table 7: Regression results for give/get verbs

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.605	0.310	5.171	0.00000
Age	0.615	0.064	9.548	0
VerbType	0.100	0.623	0.161	0.872
LogFrequency	0.985	0.392	2.512	0.012
Age:VerbType	0.728	0.141	5.157	0.00000
Age:LogFrequency	0.044	0.076	0.575	0.565
VerbType:LogFrequency	0.292	0.784	0.372	0.710
Age:VerbType:LogFrequency	-0.344	0.153	-2.253	0.024

much less improvement for three of the get-type verbs and a substantial decline in performance for *receive*.

Fig. 6 provides some context for the three-way interaction between age, log frequency, and verb type: There is initially an advantage for get-type verbs – particularly low-frequency ones – but that dissipates by ages 4 and 5 and may even start to reverse by age 6. These qualitative observations were largely confirmed by quantitative analysis. For 3 year-olds, the main effect of log frequency was significant (B = 0.92, SE = 0.30, Wald's z = 3.10, p=0.002) and the main effect of verb type trended towards significance (B = -0.86, SE = 0.49, Wald's z = -1.80, p=0.077), though their interaction was not significant (B = 0.91, SE = 0.60, Wald's z = 1.50, p=0.13). At ages 4 and 5, the only effects to even approach significance were the significant effects of log frequency (4 year-olds: B = 1.10, SE = 0.47, Wald's z = 2.40, p=0.017; 5 year-olds: B = 1.10, SE = 0.50, Wald's z = 2.20, p=0.028). At the age of 6, no effects or interactions were significant (ps $\geq$ 0.14, SE = 0.95, Wald's z = 0.42, p=0.67).

Discussion

As with psych verbs and chase/flee verbs, the give/get verbs showed a clear effect of verb type. However, it was in many ways the opposite of what was observed for the first two perspective pairs: rather than an advantage for one verb-type emerging between the ages of 3 and 6, we observed an early advantage for get-type verbs that dissipated.

It may not be necessary to make much of this difference. Three year-olds showed much higher accuracy on give-type and especially get-type verbs than they did on either type of psych verbs or chase/flee verbs – something that is consistent with their relatively early emergence in spontaneous speech (Bowerman, 1990) and extremely high token frequency. Get-type verbs must have diverged from give-type verbs at an earlier age, and had we tested two year-olds, our results may have looked (more) qualitatively similar to

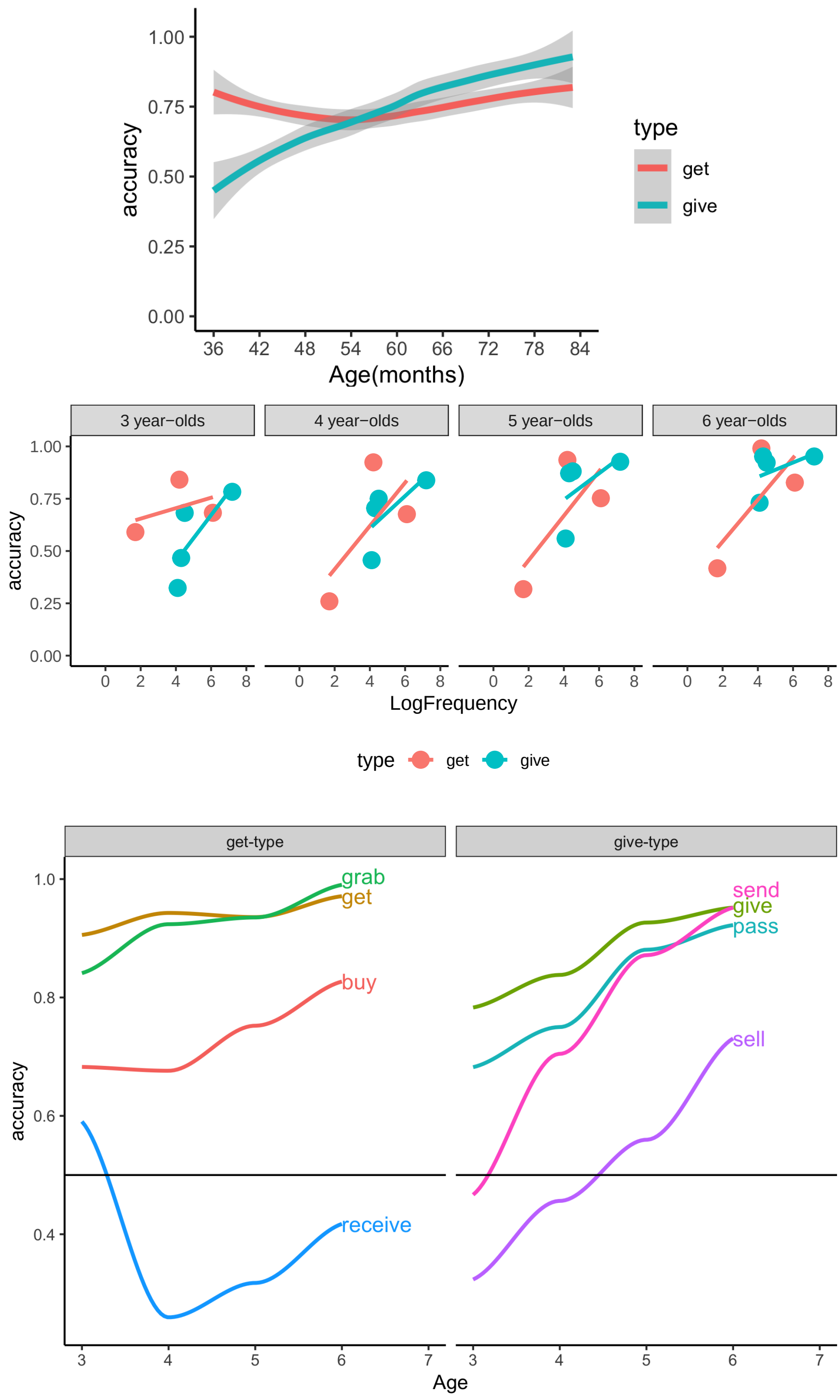

Figure 6. ***Results from Exp. 3. Top: Averaging across verbs within type, with LOESS smoothing over age. Middle: Accuracy for each verb against log frequency, split into four age groups, with linear regressions shown. Bottom: Performance on each verb across ages, aggregated into four age groups, with LOESS smoothing.***

those of the other two perspective pairs, simply shifted earlier in development.

Regardless, the findings again contradict the type-frequency hypothesis and the privileged link hypotheses, both of which predicted an advantage for give-type verbs. The privileged link hypothesis fails even if we take into account variation in the causal semantics of get-type verbs, the only class to show much variation (see Figure 7, right).

The results are, however, broadly consistent with the salience hypothesis. However, there are the puzzling U-shaped results for *receive*. Given the very large number of participants, statistical fluke seems unlikely. One possibility is that this reflects a misanalysis of the meaning of *receive*, much like what we observed for some fear-type verbs. Indeed, of all verbs in the dataset, participants in our rating study were most likely to rate *receive*'s object as causally responsible (89). Under the privileged link hypothesis, this could result in learners mistakenly treating *receive* as synonymous with *give*. Unfortunately, the rest of our results are not kind to the privileged link hypothesis, leaving this finding as something of a mystery.

General discussion

We investigated the development of high-frequency verbs in three "perspective pair" classes: emotion (psych) verbs, chase/flee verbs, and give/get verbs. In each case, argument structure type was predictive of learning, above and beyond token frequency. Taking token frequency into account, we found unexpectedly early acquisition of frighten-type verbs relative to fear-type verbs, of flee-type verbs relative to chase-type verbs, and of get-type verbs relative to give-type verbs (though by four years of age, this last difference had dissipated). Indeed, collapsing across experiments, while there is an effect of token frequency, it is clearly modulated by large effects of verb class (Figure 7, top).

This is not to say that token frequency and argument structure class are the only predictors of order of acquisition. There is additional variability not captured by these constructions (at least, to the degree we can accurately measure either; see 'Methodological Limitations', below). However, the effect was similar in size and reliability to that of token frequency, making it unusually potent by the standards of psychology.

Similarly, we cannot be sure that the verbs' argument structure causally affected pace of acquisition or was merely correlated with something that did. The obvious ways to either explain the effect of argument structure or explain it away did not pan out. The order of verb acquisition appears to be even less affected by degree to which the verb describes an event caused by the subject. Indeed, the class of verbs with greater subject causality was learned later in two of three cases (again, excluding only the psych verbs). Collapsing across all three experiments, there is little evidence of a relationship between these variables (Figure 7, middle).

Neither is the order of acquisition much predicted by the type frequency of its argument structure class. This was clear in the analyses above, where the class with greater type frequency was actually learned less well in two out of three cases (all except psych verbs). The lack of a relationship between type frequency and acquisition is made even more clear in the summary figure, which collapses across experiments (Figure 7, bottom).

The salience hypothesis is least wrong, primarily by virtue of not making many clear predictions, at least at present. It makes no predictions about fear/frighten verbs, since we currently do not know which perspective is more salient. It correctly predicts the earlier learning of get-type verbs – at least, if one adopts our interpretation of prior work on goal salience (see above). As we reviewed in the Introduction, there is currently some suspicion that the "chase" perspective is more salient than the "flee" perspective, which would be inconsistent with our finding of early learning of flee-type verbs. However, that suspicion is based on sparse data with inconsistent results. So one could reasonably argue that we have no predictions about chase/flee verbs, either.

Interestingly, all three experiments showed some evidence of U-shaped learning. U-shaped learning could be interpreted as a verb being mislearned as belonging to the wrong class. Thus, if U-shaped learning was specific to the classes that were learned later, it would be evidence of the earlier-learned classes being in some way more salient or easier to apply. However, the pattern was unclear. While four fear-type verbs showed U-shaped or below-chance learning (*trust, dislike, enjoy, fear*) – thus replicating and extending earlier observations by Hartshorne, Pogue, and Snedeker (2015) – so did two frighten-type verbs (*anger, amaze*). Among chase/flee verbs, one of each type showed below-chance learning (*pursue, flee*). Among give/get verbs, there was only one example of U-shaped learning, but it was a member of the get-type class (*receive*), which was overall acquired earlier.[14]

Methodological limitations

Before further discussing the theoretical implications, we consider several limitations in the data and the strength of the evidence.

Most broadly, we only investigated the acquisition of 30 verbs across three perspective pair classes in a single language. This is certainly a substantial improvement on prior work, which focused on narrow age ranges and dealt primarily with fear/frighten verbs in English (Bowerman, 1990; Braine et al., 1993; Hartshorne et al., 2015; Messenger et

[14]We also considered the possibility that these were the verbs for which the object was relatively causal. Indeed, the verb with strongest causality ratings for the object was *receive* (89). However, *pursue* was among those with the least causal ratings for the object (0). On the whole, there was no clear relationship.

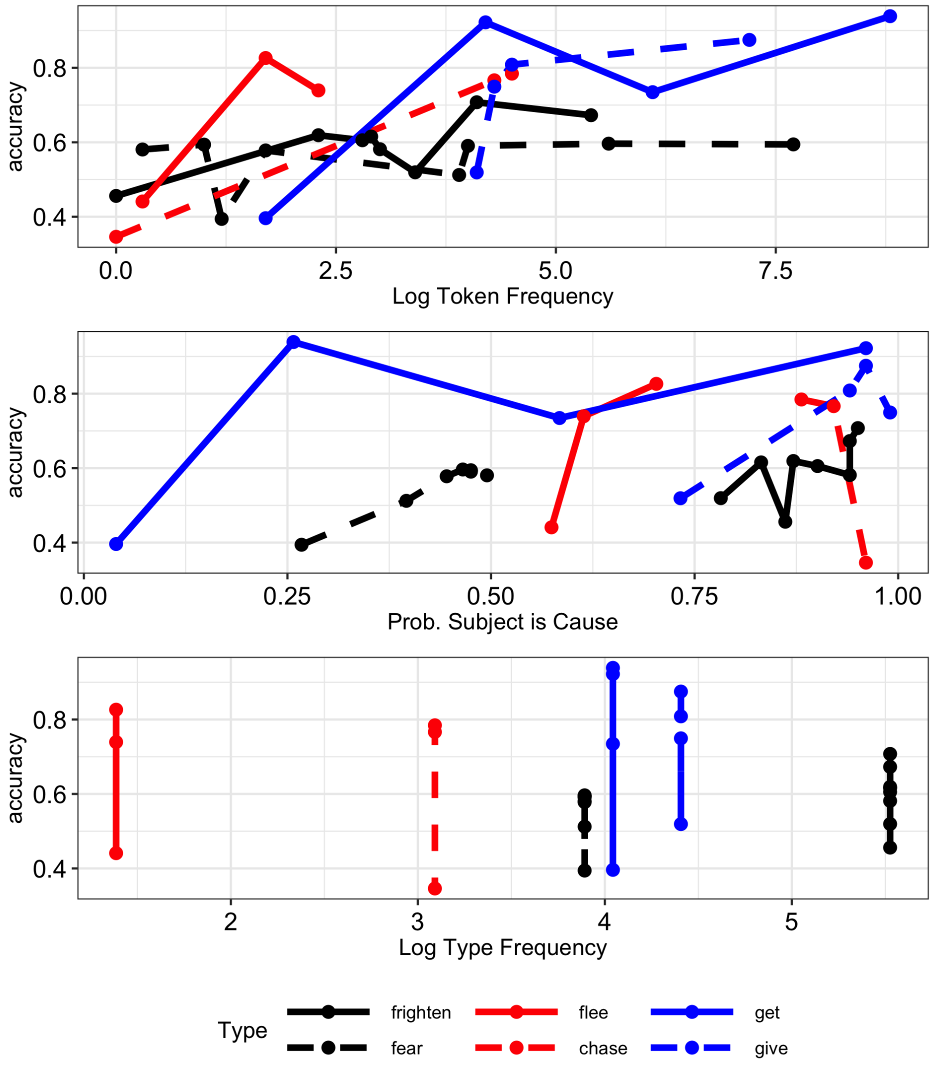

Figure 7. ***Quantitative results, summarizing across experiments. Top: Likelihood of correctly understanding a verb's argument structure is significantly if only moderately predicted by token frequency (b = 0.10, 95% CI = [0.05, 0.14], t(28) = 4.26, p < .001, R^2 = 0.39). Center: The degree to which the subject of the verb causes the event predicted accuracy even less well, an effect that approached but did not reach significance (b = 0.05, 95% CI = [-0.01, 0.11], t(28) = 1.83, p = .078; R^2 = 0.11). Bottom: There was no evidence of a relationship between accuracy and class type frequency (b = -0.02, 95% CI = [-0.08, 0.04], t(28) = -0.61, p = .545; R^2 = 0.01). Note: To facilitate comparison, the above regression coefficients are standardized.***

al., 2012; Tinker et al., 1988), with the exception of one study of the emergence in the spontaneous speech of two children of four give/get verbs (Bowerman, 1990). A few additional studies looked at relative preference for one perspective pair or the other in elicited naming, with mixed results (Fisher et al., 1994; Gleitman et al., 2007; e.g., Lakusta & Landau, 2005). While our studies included most of the verbs in the three perspective pair classes that were sufficiently high frequency in child-directed speech to be plausibly known by young children, it is quite possible the results would be different if other verbs had been available. Moreover, the results of these three perspective pairs may be unrepresentative even within English, and may not generalize to other languages – particularly languages that organize argument structure differently, such as ergative or agglutinative languages. It would certainly not be the first time such generalization failed (Evans & Levinson, 2009; Yarkoni, 2019).

Related to that point, while the sample sizes are not small by the standards of language development research, they are nonetheless not large by the standards of statistical analysis, particularly when investigating three-way interactions (Hartshorne & Schachner, 2012; Vankov, Bowers, & Munafò, 2014). Some comfort is given by the fact that in all three experiments, the key interactions that were significant in our frequentist analyses were also significant or trended towards significance in the more conservative Bayesian analyses. (The lone exception was the interaction of argument structure class and log frequency in Exp. 1.) However, interpretation of the p-values we report must take into account the fact that the small number of verbs precluded fitting maximal random effects, which may or may not be anti-conservative (Barr et al., 2013; Bates et al., 2015; Matuschek et al., 2017).

Similarly, while the estimates of frequency in child-directed speech are based on the largest dataset available (all part-of-speech-tagged corpora involving native English-speaking, North American children in CHILDES-db in the focused age range) the resulting dataset is not actually very large – around 3 million tokens, which is far less than what one child hears in a single year (Hart & Risley, 1995) and is likely to be biased by the sampling strategies used by the researchers. Moreover, many of the part-of-speech tags are automatically generated and of variable quality. Additionally, with the exception of *run from*, we counted all uses of the verbs, rather than uses in the critical syntactic frames. This was driven by a practical consideration (automatically extracting syntactic frames is difficult, particularly for spoken corpora), but determining whether it is reasonable theoretically will require more exact theories of language acquisition than we currently have.

The three experiments use three different methods: Truth Value Judgment, act-out, and question-answering. These differences were driven by the semantics of the verbs, which rendered different methods more or less natural. Consider, for instance, how one would run an act-out task for fear-type verbs, which describe a habitual disposition, not any particular action. Nonetheless, this methodological variation limits direct cross-study comparison.

Relatedly, the chase/flee experiment has a potential confound in that flee-type verbs have prepositions but chase-type verbs do not. In principle, children might have responded by ignoring the verb and focusing on the preposition. This can, of course, cut both ways: the particular sentence frame used by chase-type verbs is relatively rare compared to the transitive, which may make it more difficult for children to comprehend. This consideration just further highlights the need for data on more different kinds of perspective pair verbs.

Our test of the 'privileged link' hypothesis is contingent on our operationalization of causality. Specifically, we asked participants whether each event participant "made" the event happen. There are other ways to operationalize causality that might lead to different results. For instance, Hartshorne et al. (2016) embedded judgments of causality in a legal context, which they found resulted in particularly sharp judgments. On the other end, some authors have argued that causality is too narrow a category, preferring broader notions such as "acting on" (MacWhinney, 1977). (An anonymous reviewer has suggested we consider "actively doing something to.") More generally, the definition of "cause" is contentious even outside the linguistic domain (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021). It is unclear whether a different operationalization or causality or the selection of a different cause-related construct such as "act on" would have resulted in substantially different findings; presumably, all these constructs are reasonably correlated. Nonetheless, it will be impossible to completely rule out this worry at least until we have a full understanding of semantic representations.

Finally, as noted in the Introduction, the present studies test whether participants understood who did to whom. They do not directly address *what* was done. It is possible that participants succeeded on these tasks without understanding the differences between *love* and *hate* or between *grab* and *get*.

Theoretical implications

With the above caveats, the clearest finding is that it is not clear what drives successful verb learning. Learning was not well-predicted by token frequency, causality, or type frequency of the verb's argument class. The best that can be said is that the salience hypothesis was not entirely disconfirmed.

Thus, one obvious direction for future research is to obtain a clearer understanding of which perspective is more salient for different perspective pairs, in order to better test the hypothesis. This might be aided by developing a more nuanced, precise version of the hypothesis. Salience is a phenomenon that requires its own explanation. Perhaps a perspective is more salient because it is more simply represented (chasing involves a simple goal of being where the target is, whereas the goal of fleeing involves a negation:

not being where the pursuer is), because it is more temporally concentrated (frightening happens at a distinct time and place, whereas fearing is an ongoing state of affairs), or because of a recency effect (*get* highlights the end of an event, whereas *give* highlights its beginning; cf. Regier and Zheng (2007)). A more precise account would allow us to make predictions without necessarily having to obtain direct evidence of salience (i.e., figuring out how to measure the salience of hard-to-depict event perspectives like FEAR). For instance, researchers working on the psychophysics of action perception found it fairly straightforward to make artificial agents that chase (a simple heat-seeking policy works quite well), whereas designing artificial agents that could flee effectively required a much more complex policy (Tang et al., 2021). While this actually makes the wrong prediction in the present study (chase-type verbs were learned later, not earlier), a well-defined, quantitative simplicity-based theory of salience is potentially within reach.

Another direction would be to better characterize the quality of learning opportunities. Our findings are based on acquiring verbs earlier or later than might be expected based on input frequency. This implicitly assumes that all encounters with a verb are equally informative, which is not the case. As we reviewed above, Medina and colleagues (2011) argue that word learning is primarily driven by the rare highly informative encounter. While they present a method for identifying these 'eureka' moments, it requires time-intensive hand annotation. Currently-available annotated corpora are vanishingly small; developing one large enough to test whether frighten-type, flee-type, and get-type verbs have more than their fair share of eureka moments will either require an enormous amount of work or some mechanism for automating the annotation (e.g., through machine learning). Note, moreover, Hartshorne, Pogue, and Snedeker (2015) raise some reasons for being skeptical about this explanation, at least with respect to fear/frighten verbs. They suspect that because frighten-type events are ephemeral, speakers are unlikely to remark upon them as they happen, whereas because fear-type states are ongoing, the reverse may be true for them. As a result, it may be easier for children to connect a fear-type utterance with its co-temperal referent than a frighten-type utterance with its non-co-temporal referent.

Encounters with verbs can be more or less informative in other ways as well. Hartshorne, Pogue, and Snedeker (2015) reported that high-frequency fear-type verbs such as *like*, *love*, and *hate* occur primarily with one of a small number of subjects – mostly *I* and *you*. They note that this might induce children to treat these high-frequency bigrams as set phrases. As a result, most uses of these verbs would fail to provide much information about their argument structure. In that case, the frequency analyses above should be redone excluding those high-frequency bigrams. This is not trivial – for one thing, it requires a principled way of determining which bigrams are of sufficiently high frequency – and we leave it to future work.

More generally, verbs differ in many ways beyond argument structure class, token frequency, type frequency, causal structure, and salience. As already noted, an anonymous reviewer suggests that in our data, learning seems to be somewhat earlier for verbs involving events where one or more participants are particularly active and intentional. One can certainly imagine this makes the events more cognitively salient or simply easier to spot in the world. Other cognitive biases, such as a bias towards positive (or negative) events could also play a role.[15] There are currently a vast range of possibilities to explore, given that the ones most grounded in the literature are less explanatory than anticipated.

Conclusion

Hartshorne, Pogue, and Snedeker (2015) reported a puzzling finding: relatively old children failed to understand extremely high-frequency fear-type verbs long after they had acquired a number of lower-frequency frighten-type verbs. They proposed a number of possible explanations based on current understanding of verb-learning. Of the ones that are currently testable, none are clearly consistent with the present results. More broadly, the current study revealed that this is not a funny fact about fear/frighten verbs, but may in fact be a common phenomenon – one that has gone largely undetected and remains essentially without explanation. This conclusion is based on only three case studies in a single language: the empirical picture may be even more complicated than it appears so far. All we can say at the moment is that something is going on, and we do not understand it. This should concern us, because if we are missing a large part of the empirical description of language acquisition, our theorizing may be entirely misdirected. At the very least, it is incomplete. There is more in heaven and earth than is dreamt of in our philosophy, so some new philosophy is needed.

[15]We thank an anonymous reviewer for this specific suggestion.

References

Ambridge, B., Barak, L., Wonnacott, E., Bannard, C., Sala, G., Zwaan, R., & Ferreira, F. (2018). Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis. *Collabra: Psychology*, *4*(1).

Ambridge, B., Maitreyee, R., Tatsumi, T., Doherty, L., Zicherman, S., Pedro, P. M., et al.others. (2020). The crosslinguistic acquisition of sentence structure: Computational modeling and grammaticality judgments from adult and child speakers of english, japanese, hindi, hebrew and k'iche'. *Cognition*, *202*, 104310.

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv Preprint arXiv:1506.04967*.

Belletti, A., & Rizzi, L. (1988). Psych-verbs and θ-theory. *Natural Language & Linguistic Theory*, 291–352.

Bowerman, M. (1990). Mapping thematic roles onto syntactic functions: Are children helped by innate linking rules? *Linguistics*, *28*(6), 1253–1289.

Braine, M. D. (1992). What sort of innate structure is needed to "bootstrap" into syntax? *Cognition*, *45*(1), 77–100.

Braine, M. D., Brooks, P. J., Cowan, N., Samuels, M. C., & Tamis-LeMonda, C. (1993). The development of categories at the semantics/syntax interface. *Cognitive Development*, *8*(4), 465–494.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychome-*

trika, 78(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Dowty, D. R. (1989). On the semantic content of the notion of “thematic role.” In *Properties, types and meaning* (pp. 69–129). Springer.

Dowty, D. R. (1991). Thematic proto-roles and argument selection. *Language, 67*(3), 547–619.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*(5), 429–448.

Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(2), 143–149.

Fisher, C., Hall, D. G., Rakowitz, S., & Gleitman, L. R. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua, 92*, 333–375.

Freeman, N. H., Sinha, C. G., & Stedmon, J. A. (1981). The allative bias in three-year-olds is almost proof against task naturalness. *Journal of Child Language, 8*(2), 283–296.

Fujita, H. I. I. (2000). A cognitive approach to errors in case marking in japanese agrammatism. *Constructions in Cognitive Linguistics: Selected Papers from the Fifth International Cognitive Linguistics Conference, Amsterdam, 1997, 178*, 123. John Benjamins Publishing.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; No. 257.*

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review, 128*(5), 936.

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition, 1*(1), 3–55.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language, 57*(4), 544–569.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language.*

Oxford University Press on Demand.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35*(3), 515–531.

Grimshaw, J. (1990). *Argument structure.* the MIT Press.

Hansen, P. (2017). What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development. *First Language, 37*(2), 205–225.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Paul H Brookes Publishing.

Hartshorne, J. K., Bonial, C., & Palmer, M. (2014). The VerbCorner project: Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 397–402.

Hartshorne, J. K., O'Donnell, T. J., Sudo, Y., Uruwashi, M., Lee, M., & Snedeker, J. (2016). Psych verbs, the linking problem, and the acquisition of language. *Cognition, 157,* 268–288.

Hartshorne, J. K., Pogue, A., & Snedeker, J. (2015). Love is hard to understand: The relationship between transitivity and caused events in the acquisition of emotion verbs. *Journal of Child Language, 42*(3), 467.

Hartshorne, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience, 6,* 8.

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables.* Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). Retrieved from https://CRAN.R-project.org/package=stargazer

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of english verbs. *Language Resources and Evaluation, 42*(1), 21–40.

Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. *Cognition, 96*(1), 1–33.

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science, 36*(3), 517–544.

Landau, B., & Gleitman, L. R. (2015). 10 height matters. *Structures in the Mind: Essays on Language, Music, and Cognition in Honor of Ray Jackendoff*, 187.

Levin, B., & Hovav, M. R. (2005). *Argument realization*.

Lidz, J., Gleitman, H., & Gleitman, L. R. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, *87*(3), 151–178.

Lüdecke, D. (2018). Ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772

MacWhinney, B. (1977). Starting points. *Language*, 152–168.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Marantz, A. (1982). On the acquisition of grammatical relations in linguistik als kognitive wissenschaft. *Linguistische Berichte Braunschweig*, (80), 32–69.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type i error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M., & Lannon, R. (2011). An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental Science*, *14*(2), 181–189.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Messenger, K., Branigan, H. P., McLean, J. F., & Sorace, A. (2012). Is young children's passive syntax semantically constrained? Evidence from syntactic priming. *Journal of Memory and Language*, *66*(4), 568–587.

Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, *34*(6), 1064–1092.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, Mass.: Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.*

Quine, W. (1964). *Word and object.*

Regier, T., & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science, 31*(4), 705–719.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods, 51*(4), 1928–1941.

Strickland, B., Fisher, M., Keil, F., & Knobe, J. (2014). Syntax and intentionality: An automatic link between language and theory-of-mind. *Cognition, 133*(1), 249–261.

Tang, N., Gong, S., Liao, Z., Xu, H., Zhou, J., Shen, M., & Gao, T. (2021). Jointly perceiving physics and mind: Motion, force and intention. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*.

Tinker, E., Beckwith, R., & Dougherty, R. (1988). Markedness and the acquisition of emotion verbs.

Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article commentary: On the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology, 67*(5), 1037–1040.

Wellman, H. M. (1992). *The child's theory of mind.* The MIT Press.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Yarkoni, T. (2019). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37.

Data, Code and Materials Availability Statement

Data, code, and materials are available at https://osf.io/k5xud/.

Ethics statement

Ethics approval was obtained from the ethics committee of the University of Nowhere. Parents of child participants gave gave informed written consent before the child took part in the study. Adult participants recruited and tested online read a consent statement before participating.

Authorship and Contributorship Statement

JKH conceived of the study. JKH and LS designed the experiments. JKH, YH, and LS performed analyses. JKH and YH wrote the manuscript.

Acknowledgements

The authors thank Jesse Snedeker, Tim O'Donnell, Laura Lakusta, Vera Kempe, and two anonymous reviewers for helpful feedback and suggestions, and Miguel Mejia, Zach Barker, Kayley Okst, Caitlin Garcia, Nicola Roux, Marissa Russell, Camille Phaneuf, Ernesto Gutierrez, Juliani Vidal, Hayley Greenough, Casey Nicastri, Rudmila Rashid, Taylor Martinz, Kate Roberts, Madeleine McCanne, Sarah Al-Mayahi, Alice Lim, Everett Kim, Lily Feinberg, Zeen Naeem and the staff at the Acton Discovery Museum and Boston Children's Museum for assistance with data-collection. Funding was provided by NSF 1551834, awarded to JKH.