

**Semantic cues in language learning:**

**An artificial language study with adult and child learners**

Helen Brown<sup>a</sup>, Kenny Smith<sup>b</sup>, Anna Samara<sup>c</sup>, & Elizabeth Wonnacott<sup>d</sup>

*<sup>a</sup> Psychology Department, Nottingham Trent University, UK*

*<sup>b</sup> School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, UK*

*<sup>c</sup> School of Human Sciences, University of Greenwich, UK*

*<sup>d</sup> Division of Psychology and Language Sciences, University College London, UK*

Correspondence concerning this article should be addressed to Helen Brown, Psychology Department, Nottingham Trent University, Burton Street, Nottingham, NG1 4BU, UK.

Email: [helen.brown@ntu.ac.uk](mailto:helen.brown@ntu.ac.uk)

Declarations of interest: None

### **Abstract**

Grammatical regularities may correlate with semantics; e.g., grammatical gender is often partially predictable from the noun's semantics. We explore whether learners generalise over semantic cues, and whether extent of exposure (1 versus 4 sessions) and number of exemplars for each semantic class (type-frequency) affect this. Six-year-olds and adults were exposed to semi-artificial languages where nouns co-occurred with novel particles, with particle usage fully or partially determined by the semantics of nouns. Both adults and children generalised to novel nouns when semantic cues were fully consistent. Adults (but not children) also generalised when cues were partially consistent. Generalisation increased with exposure, however there was no evidence that increasing type-frequency (i.e. more nouns per semantic class) increased generalisation. Post-experiment interviews also suggested that successful generalisation depended on explicit awareness. These results suggest that semantic cues are particularly difficult for children to exploit during the early stages of language acquisition.

**Key words:** Artificial language learning; language acquisition; statistical learning; semantic cues

## Introduction

Language acquisition involves generalisation, which enables us to use words in novel, yet grammatically well-formed ways. Generalisation may occur on the basis of distribution, but also similarities between words in terms of phonological or semantic properties. For example, many languages use grammatical gender, where subgroups of nouns are associated with different grammatical forms. Grammatical gender may appear arbitrary (Bloomfield 1933; Maratsos, 1982), yet corpus analysis reveals surprising regularities (e.g., Corbett, 1991; Mirković, MacDonald, & Seidenberg, 2005): in Serbian, nouns referring to fruits tend to be feminine, while nouns referring to vegetables tend to be masculine (a semantic regularity); in French, words ending in *-ette* are more likely to be feminine than masculine (a phonological regularity). Adult native speakers are sensitive to such regularities, as evidenced by their usage of gender markers with novel words (Arias-Trejo & Alva, 2013; Karmiloff-Smith, 1981; Mulford, 1985), and naturally occurring speech errors (Barbaud, Ducharme, & Valois, 1982; Szagun, Stumper, Sondag, & Franik, 2007; Vigliocco, Vinson, Martin, & Garrett, 1999). A key question for language acquisition is whether, and under what circumstances, children also make such generalisations.

One approach to this question looks at children's sensitivity to cues within natural languages. At least for gender, some evidence suggests that child learners disproportionately favour phonological cues when they occur with semantic cues. For instance, until age 10, native French-speaking children use phonological rather than semantic cues to determine their usage of gender marked forms with novel nouns, despite the fact that natural gender (e.g., the biological gender of a person, animal, or character, which is a semantic cue) strongly predicts gender class in French (Gagliardi & Lidz, 2014; Karmiloff-Smith, 1981; Pérez-Pereira, 1991). One exception is Mulford (1985), who showed that Icelandic children were sensitive to natural gender from age 4, whilst phonological cues only affected

generalisation in older children (7-8-year-olds). Mulford attributes this to the unreliable and often multifunctional nature of phonological cues in Icelandic, which suggests that the consistency and robustness of cues in the input may be critical. Alternatively, young children may rely more on phonological cues than semantic cues simply because of their earlier availability (i.e., infants are exposed to word forms before they successfully acquire form-meaning mappings; Gagliardi, Feldman, & Lidz, 2017). These two explanations are consistent with Bates and MacWhinney's (1989) competition model of syntactic processing. This model assumes a key role for cue validity, which refers to both cue availability (how often the cue is present in the input) and cue reliability (how often the cue leads to the correct parsing of the syntax) (Li & MacWhinney, 2012). A final possibility is that children disprefer semantic cues for reasons unrelated to their relative consistency or availability. For example, Ferman, Olshtain, Schechtman and Karni (2009) suggest that exploiting semantic cues relies more on declarative memory than phonological learning, and declarative memory is known to show strong effects of age-dependent maturation (Digiulio, Seidenberg, O'Leary, & Raz, 1994). However, importantly, children are clearly capable of learning how semantic properties link to lexical items. For example, they may know how natural gender relates to the superordinate terms for *man* and *woman*. It is relating these cues to grammatical gender that appears to be difficult.

In studies of natural language learning, where semantic and phonological cues are confounded, it is difficult to distinguish between these possibilities. Artificial language learning paradigms, where learners are exposed to experimenter-created languages, provide an alternative methodology for exploring generalisation over particular types of cues. Early work demonstrated that, although in principle word distribution provides evidence that words fall into grammatical categories (Cartwright & Brent, 1997; Mintz, Newport, & Bever, 2002; Redington, Chater, & Finch, 1998), participants (primarily adults) only abstracted category

generalisations from the input when there were correlated phonological cues (Aurilio, Jenkins & Silliman, 2000; Braine et al., 1990; Brooks et al., 1993; Frigo & McDonald, 1998; Gerken, Wilson, & Lewis, 2005; Smith, 1969), semantic cues (Braine, 1987; Ferman et al., 2009; Leung & Williams, 2012; Williams, 2005), or convergent phonological and semantic cues (Mirković, Forrest, & Gaskell, 2011) (though see Mintz, Wang, & Li, 2014). This is in-line with phonological/semantic bootstrapping accounts, whereby external cues play a key role in the early stages of abstracting grammatical categories from the input (Grimshaw, 1981; Morgan & Demuth, 2014; Pinker, 1984).

Other work with artificial languages explores how input structure influences the interplay between different types of cues (e.g. Monaghan, Chater & Christiansen, 2005). Culbertson, Gagliardi, and Smith (2017) looked at the use of phonological and semantic cues by adult learners and manipulated the order in which these became available in exposure to an artificial language. They found greater reliance on early learned cues, whether phonological *or* semantic. This occurred even when later learned cues were more salient This preference for early learned cues supports the early availability explanation for children's preference for phonological over semantic cues (e.g., Karmiloff-Smith, 1981). Thus, adult learners' propensity to exploit semantic cues may depend on input structure.

In contrast, relatively few relevant artificial language studies have looked at children. Some of these explore how generalisation is affected by word distribution (e.g. Reeder, Newport & Aslin, 2017; Saffran, 2001; Wonnacott, 2011; Wonnacott, Brown, & Nation, 2017). For semantic and phonological cues, while there is clear evidence that children, like adults, can generalise on the basis of isolated phonological cues (e.g., Brooks et al., 1993, Gerken et al., 2005), the evidence is more mixed for semantic cues in isolation. Ferman and Karni (2010) found that 12-year-olds and adults, but not 8-year-olds generalised a novel morphological rule to new nouns using a semantic cue (animacy). However, in the majority

of cases, those who generalised were explicitly aware of the animacy decision. This was taken to show that learning of the novel language rules depended on explicit memory, with explicit learning relying on declarative memory systems (Ullman, 2016) that develop relatively slowly through childhood (Digiulio et al., 1994), although other adult studies have shown generalisation over semantic cues without explicit awareness (Leung & Williams, 2012; Vujovic, Ramscar & Wonnacott, under review; Williams, 2005), a point we return to in the Discussion.

Similarly, Schwab, Lew-Williams and Goldberg (2018) found that adults, but not six-year-olds, could generalise a novel classifier to new nouns on the basis of a semantic cue (natural gender) following training with a partially consistent semantic cue (three of the four training items embodied natural male or female gender, and the fourth item was an inanimate object). Together, Ferman and Karni (2010) and Schwab et al.'s (2018) results suggest that children's difficulty in using semantic cues in natural language learning may not stem solely from the fact that semantic cues are available later in learning, nor from a general inability to learn cues to noun class (because phonological cues are learnable); rather, semantic cues may be particularly difficult for young children to access. In contrast to these studies, Lany and Saffran (2010; see also Lany & Saffran, 2011) found that 22-month-old infants demonstrated generalisation over a fully consistent semantic cue (they learned that animals occurred with one determiner and vehicles with another). However, in this study, word classes were also marked with converging phonological cues. To our knowledge, only one published study with children has demonstrated generalisation over semantics alone. Using an artificial language, Culbertson, Jarvinen, Haggarty, and Smith (2019, Experiment 1) demonstrated that when semantic cues were available for every noun and *consistently* predicted the co-occurring determiner 6-7-year-olds generalise novel nouns to appropriate determiners based on their semantic features.

Why did Culbertson et al. (2019) find generalisation over semantic cues, unlike Schwab et al. (2018) and Ferman and Karni (2010)? Firstly, Culbertson et al. used fully consistent semantic cues, whilst Schwab et al. (2018) included some exception nouns as is common in natural languages (e.g. in German, e.g. Mädchen, girl, is neuter rather than feminine). An artificial language study by Samara, Smith, Brown, and Wonnacott, (2017) found that partially reliable (here, social) cues were harder to learn, particularly for children, with evidence of learning only after four training lab sessions, even in adults. In comparison, a single training session sufficed for learning of fully consistent conditioning in Culbertson et al. (2019). On the other hand, conditioning in Ferman and Karni's (2010) study *was* fully consistent, yet their participants didn't show generalisation even after ten sessions of training. However, their participants were speakers of a language containing gender classes (Hebrew), who then had to learn *new* gender categories, with different cues from their existing categories. Thus their previous knowledge may perhaps have interfered with learning in this study.

Another factor which may have decreased the likelihood of generalisation in previous artificial language studies is *type-frequency* i.e., the number of unique nouns exemplifying the semantic cues during exposure. Ferman and Karni (2010) used 16 noun-verb exemplars, but do not report the number of unique nouns involved. Schwab et al. (2018) had only three lexical items exemplifying each type of semantic cue. Previous research suggests that encountering variable exemplars promotes generalisation (e.g. Bybee, 1995; Gomez, 2002; Plunkett & Marchman, 1991, 1993; Wonnacott, Boyd, Thomson, & Goldberg, 2012). This is in line with theoretical approaches in which generalisation is a probabilistic process that involves distinguishing the relevant cues (e.g. "animalness") from irrelevant cues (e.g. idiosyncratic features associated with particular animals) (Ramscar et al., 2010; Apfelbaum & McMurray, 2011).

Taken together, studies that exploit different types of cues (summarized in Appendix A) strongly suggest that children, like adults, can generalise words to novel grammatical contexts based on phonological cues, even when they are only partially predictive or the only available cue. In contrast, there is limited evidence that this is true for semantic cues, particularly when their usage is not fully consistent. These findings suggest that semantic cues are particularly difficult for young children to access and use during the early stages of language acquisition. However, methodological limitations of previous studies (interference from Hebrew for Ferman and Karni, 2010; low type-frequency and/or insufficient exposure for Schwab et al., 2018) may account for at least some of these difficulties. In the current work, we further explore children's (and adult's) ability to generalise over semantic cues. Our aims were to replicate the finding of Culbertson et al. (2019) that children can generalise across fully consistent semantic cues, and to determine whether children can also generalise across *partially* consistent semantic cues (which is more representative of such cues in natural languages) given sufficient exposure / increased type-frequency. Our approach is to explore learning of semantic cues in isolation, since this is a prerequisite to understanding learning when multiple cues co-occur, as they inevitably do in natural languages, and compete or undergo integration.

### **The Current Study**

We employed a multi-session semi-artificial language learning paradigm (similar to Samara et al., 2017) to explore the learning of semantic cues by 6-year-olds and adults. Participants were monolingual native English speakers, with limited or no experience of languages containing gender classes. We tested 6-year-olds for two reasons. First, as can be seen in Appendix A, previous studies looking at gender in natural language suggest that 6-year-olds are roughly in the middle of the age range shown to differ from adults and older children (age 10+) in their propensity to use semantic cues. Second, Culbertson et al. (2019)



showed that 6-year-olds privilege phonological over semantic cues in controlled competing conditions. Thus, 6-year-olds are a suitable model population for younger learners, at least with respect to generalisation over semantic cues.

Training occurred over four sessions, addressing the concern that Schwab et al. (2018) did not provide sufficient exposure to the novel language. As in Samara et al. (2017), we manipulated *cue consistency*, contrasting the learning of languages with *fully consistent* and *partially consistent* cues. Nouns in the target language referred to animals or vehicles and were followed by one of two novel particles (e.g. *dak* and *pag*) which occurred equally frequently in their input. In the fully consistent condition, the semantic category of the noun perfectly predicted particle choice. We chose a semantic cue (animals vs vehicles) that should be well-known to children of this age: our interest was not in learning a new semantic distinction, but rather whether children could leverage an existing semantic distinction as a cue to facilitate learning of a new grammatical distinction. This should provide the most conducive circumstances for learning a semantic cue to noun class, in that it involves mapping a salient pre-existing semantic cue to a particle; in natural language learning children might have to simultaneously learn the salient semantic feature. In the partially consistent condition, each noun class contained one exception noun, which, unlike Schwab et al. (2018), had the semantic feature of the *other* noun class (i.e. if other nouns referring to animals were followed by particle 1, one exception noun referring to a vehicle occurred with particle 1). We also included an *inconsistent* condition whereby half of the items from each semantic class appeared with both particles – i.e. semantics did not cue particle choice. This condition allowed us to explore whether item-based learning affects higher-level generalisations over semantics (Perfors, Tenenbaum, & Wonnacott, 2010): Do participants in the inconsistent language (where there are no “helpful” semantic cues which might boost learning) remember trained noun-particle associations as successfully as participants exposed

to languages containing semantic cues? Finally, we manipulated *type-frequency* – the number of unique nouns exemplifying the semantic cues in the training set – across each of the consistency conditions. This addresses the concern that some of the previous studies may not have provided sufficiently varied input for generalisations over semantic cues to occur.

We predicted: generally stronger learning in adults than children; more successful and quicker learning when the cues were fully rather than partially consistent; stronger generalisation of semantic cues to novel items following high type-frequency input; and, focusing in particular on child learners, we predicted that generalisation to novel items was more likely to emerge in the more consistent conditions, after 4 sessions of training, and under conditions of higher type-frequency.

Finally, we explored whether awareness of the relationship between particle usage and semantic class was important given findings in some studies (Ferman & Karni, 2010; Ferman et al., 2009) that semantic generalisation depended on explicit awareness.

## Method

### Participants

Ninety 6-year-olds ( $M_{\text{age}} = 6;0$ ,  $SD = 0;5$ , 34 male) and 60 adults (university students;  $M_{\text{age}} = 19;7$ ,  $SD = 2;3$ , 13 male) participated. Fifteen children and ten adults were randomly assigned to each of the six experimental conditions (see below). Participants' scores in standardized memory tests (included for exploratory purposes to determine whether short-term or working memory affected learning) and other sample descriptives are reported in Table 1.<sup>1</sup> For children, written parental consent was obtained, as well as verbal assent before each session. Children were rewarded with stickers and a certificate. Adults provided written consent and were rewarded with partial course credit or payment. Participants were monolingual native English speakers<sup>2</sup> with no known hearing, language, or speech disorders.

[Table 1 about here]

## Stimuli

The stimulus set consisted of one sentence-initial word (*glim*), 16 English nouns denoting animals (e.g., *cow*, *dog*), 16 English nouns denoting vehicles (e.g., *ambulance*, *bike*), and ten sentence-final particles (e.g., *bup*, *dak*).<sup>3</sup> Sentences took the form *glim noun particle*, where *glim* was a carrier phrase that allowed us to prompt participants during production without providing meaningful content. Two of the ten sentence-final particles were randomly selected for each participant to minimise potential biases associated with particular particles or noun-particle pairings (e.g. alliteration – *bus bup*). Particle use was conditioned (with various levels of consistency; see below) by the animacy of the noun. More specifically, the particles differentiated animals (animate) from vehicles (inanimate), which are familiar to 22-month-olds (Lany & Saffran, 2010), and are thus suitable for 5- 6-year-olds.

Stimuli were recorded by a female British English speaker. Words were edited into separate sound files, and peak amplitude was normalised using Audacity. Clipart pictures of the 32 nouns (e.g., two tigers) were obtained online. By design, animal and vehicle words were well matched for length (in number of phonemes and syllables), frequency, and rated age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012).

## Design

As shown in Table 2, two variables were manipulated in the learning input: the semantic consistency with which particles were used during training, and type-frequency (the number of exemplars in each category).

Regarding semantic consistency, particles were either (i) *fully consistent* – particle 1 occurred only with animals and particle 2 occurred only with vehicles; (ii) *partially*

*consistent* – all-but-one animal occurred with particle 1 (with the exception animal always occurring with particle 2), and all but one vehicle occurred with particle 2 (with the exception vehicle always occurring with particle 1); (iii) *inconsistent* – half of the animals and half of the vehicles occurred with particle 1, the other half with particle 2. Note that in the inconsistent condition each noun consistently occurred with the same particle, but the semantic category of the noun was not predictive of that particle.

Regarding type-frequency, participants either received a *low type-frequency* training set of 4 animals and 4 vehicles, or a *high type-frequency* training set containing 8 animals and 8 vehicles. This allowed the semantic classes in the *fully consistent* and *partially consistent* conditions to be exemplified with more nouns in the high type-frequency than in the low type-frequency condition. In all cases, training nouns were randomly selected for each participant from the full set of 16 animals and 16 vehicles; 8 additional nouns were selected at random from each category as novel test nouns (see below). Each noun was encountered twice as often in the low frequency conditions, so that total training duration and total frequency of the novel particles was matched across conditions.

*Semantic consistency* and *type-frequency* were fully crossed, yielding six conditions in total (see Table 2) tested between subjects. Six-year-olds and adults learned and were tested on the semi-artificial languages over four sessions, with tests at the end of Sessions 1 and 4, allowing us to look (within-subjects) at the role of increased exposure. The majority of participants (82/90 children, 54/60 adults) completed four sessions on four consecutive days. The remaining participants were tested over a maximum of eight days. The tasks completed in each session are summarised in Table 3.

[Tables 2 and 3 about here]

## **Procedure**

Child and adult participants were tested individually under identical instructions. Tasks were run using ExBuilder software (a custom-built software package developed at the University of Rochester). Participants were introduced to a toy frog and were told that they were going to learn “Freddy Frog’s language”.

### ***Noun Practice***

Participants first practiced saying the names of all the nouns in their training and test sets. In the low type-frequency condition, these were 12 animals and 12 vehicles (for each semantic category four items featured in sentence training trials, and unbeknownst to participants, eight featured as novel items in production tests, four per test session). In the high type-frequency condition these were 16 animals and 16 vehicles (eight trained items and eight novel items per semantic category). In Session 1, participants completed two noun practice tasks. First, they viewed a picture of a single animal or vehicle (e.g., a tiger), heard the corresponding English word (e.g. *tiger*), and repeated the word aloud. Second, they repeated the task without prompts. The latter task was repeated at the beginning of all subsequent sessions to discourage children from using unintended labels (e.g., *bunny* for *rabbit*): these were corrected by the experimenter (e.g. *Freddy calls this one a rabbit. Can you say rabbit?*).

### ***Sentence Training***

On each trial, participants saw a picture two animals/vehicles of the same type (e.g., two tigers)<sup>4</sup>, heard a sentence (e.g., *glim tiger bup*) and repeated it aloud. Mispronunciations were corrected once. In all conditions, there were 64 training sentences, each encountered once per session. These were administered in a single block in Sessions 1 and 4, and split into two blocks of 32 trials in Sessions 2 and 3. The composition of the training set varied by condition as shown in Table 2. Note that total exposure to the particles was matched across

conditions, meaning that individual nouns were more frequent in the low type-frequency conditions.

### ***Production Test***

Following sentence training in Sessions 1 and 4, participants completed a production test. On each test trial, participants saw a picture (e.g., two tigers), heard *glim*, and were asked to finish the sentence. Incorrect noun responses were corrected (e.g., *Good try, but this one is a tiger, not a lion*) and participants were asked to say the sentence again using the correct noun. These trials were scored as incorrect.<sup>5</sup> No feedback was provided regarding sentence-final particles. If no particle was produced (e.g., *glim tiger*), children were asked if they were ready to move to the next trial.

There were 64 trials in each production test. The first eight trials always used trained nouns (four animals and four vehicles, tested once each)<sup>6</sup> and the remaining 56 trials tested performance on the eight trained nouns (seen a further three times each) alongside four novel nouns per category (animal/vehicle) that had not been encountered during training (tested four times each). Item order was pseudo-randomised, to prevent consecutive repetitions of the same noun. Identical trained items were used in Sessions 1 and 4, but different novel nouns were used at each test point.

### ***Two-Alternative Forced Choice (2AFC) Test***

Participants completed this test in Session 4 only. They were told that they would be helping Freddy's friends to say things like Freddy. On each trial, a picture (e.g., two tigers) appeared at the top of the screen. An image of a cartoon frog with a speech bubble then appeared in the bottom left corner of the screen, and participants heard sentence 1. Finally, a second frog appeared in the bottom right corner of the screen, and participants heard sentence 2. Sentences always took the form *glim noun particle 1* and *glim noun particle 2* (left/right position randomised). Participants clicked on the frog whose sentence best described the

picture. Eight trained (four animals, four vehicles) and eight novel (four animals, four vehicles) nouns (also used in the Session 4 production test) were tested, with each noun presented once each in a random order.

### ***Standardized Memory Measures***

Participants completed three standardized memory measures from the Automated Working Memory Assessment (Alloway, 2008). Verbal short-term memory was measured in Session 2 using a word recall task. Non-verbal short-term memory and verbal working memory were measured using maze memory and backwards digit recall, respectively, in Session 3.

### ***Post-Experiment Interview***

At the end of Session 4, participants were asked questions (listed in Appendix B) assessing/prompting their ability to describe any patterns they had noticed during learning. Based on their responses (i.e., whether they could describe that different particles co-occurred with different semantic categories), they were binary coded as being/not being aware of the association between particle use and the semantic categories. Links with the semantic categories could be described using either superordinate-level labels (e.g., *living; non-living*), basic level labels (e.g., *animals; vehicles; machines*), subordinate-level descriptions (e.g., *cars and things like that*) or feature-level descriptions (e.g., *things that you get in; things with wheels*). Participants who indicated awareness of the semantic cues but attributed the wrong particle to the semantic categories (two children, one adult) were scored as unaware.

## **Results and Discussion**

### **Overview of Statistical Analyses**

Data from the two tasks (production and 2AFC) were analysed separately. Separate analyses were also carried out for trained and novel nouns. For the former, we analysed data from all conditions (fully consistent, partially consistent, inconsistent) whereas, for novel

nouns (which test generalisation), we excluded the inconsistent condition (where there are no consistent semantic cues to generalise over). Except for targeted age-group comparisons, data from children and adults were also analysed separately.

Our set of first core analyses used logistic mixed effect models to explore effects of semantic consistency, type-frequency (and age-group in the age comparisons) and their interactions on performance (producing/choosing the correct determiner). We then further probed the novel noun data for any evidence of above chance generalisation when those participants who reported noticing the semantic cues in the post-experiment interview were excluded. A final set of analyses followed up on critical non-significant effects in the main analyses: specifically, Bayes Factor analyses were conducted to determine whether there was evidence to support the null hypothesis in each case (since frequentist  $p$  values do not provide this information).

Further to the analyses reported here, additional analyses comparing performance on exception versus majority-particle trained nouns in the partially consistent condition are included in Appendix C. Supplementary online analyses looking at relationships between performance and memory measures, and other analyses using measures of regularization (*cf.* Hudson Kam & Newport, 2005; Schwab et al., 2018) are also available (<https://osf.io/sy8zr/>).

### **Logistic Mixed Effect Models**

Accuracy data (correct/incorrect particle usage) were analysed by logistic mixed effects models (Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Quene & van den Bergh, 2008) using the *lme4* package (Bates, Maechler, & Bolker, 2013) in R (R Development Core Team, 2010). Data from trained and novel nouns were analysed separately for each test (production, 2AFC). The key findings from these analyses are summarised in Table 4 and detailed below.

Each model included all relevant experimentally manipulated variables and all interactions between those variables as fixed factors, regardless of whether they contributed



significantly to the model. All predicting variables (including discrete factor codings) were centred (i) to reduce collinearity between main effects and interactions; (ii) so that the main effects were evaluated as the average effects over all levels of the other predictors (rather than at a specified reference level for each factor). Effects were coded such that a positive coefficient was in the direction of the key prediction (i.e., stronger performance where consistency is greater, for high type-frequency than for low type-frequency and in session 4 than session 1). *Participant* was included as a random effect and a full random slope structure was used in each model, as recommended by Barr, Levy, Scheepers, and Tily (2013), unless otherwise stated. These models converged with Bound Optimization by Quadratic Approximation (Powell, 2009). The data, R analyses script, and model outputs are available at <https://osf.io/sy8zr/>. Effects which are not reported were not significant ( $p > .05$ ).

[Table 4 about here]

## **Production Data: Accuracy**

### ***Data Preparation***

Trials were removed if an incorrect noun was produced (children: 2.5%; adults: 0.4%); if no particle was produced (children: 0.4%; adults: 0.01%); or if the final particle was not clearly identifiable as one in the input (children: 13.2%; adults: 2.8%). Mispronunciations that resulted in identifiable particles (e.g., a single phoneme substitution such as *tib* → *tid*) were retained.<sup>7</sup>

### ***Trained Nouns***

Figure 1 shows the proportion of correctly produced particles for trained nouns in each session and provides information about whether participants were aware of (or at least able to verbally report in the post-experiment interview) the association between particle use

and semantic category. First, we analyse performance across aware and unaware participants combined in models predicting particle accuracy by (the fixed factors) *semantic consistency* (fully consistent/ partially consistent/ inconsistent), *type-frequency* (high/low), *session* (1/4), and their interactions.

[Figure 1 about here]

For semantic consistency, our predictions concerned the contrasts between three conditions rather than looking for an overall effect of semantic consistency, which is not interpretable. We thus inspected the model for these contrasts and how they interacted with the other factors using a successive differences coding of the consistency factor. This allows us to compare each level of consistency to the preceding level (fully consistent to partially consistent, partially consistent to inconsistent – coded such that a positive beta indicates stronger performance for higher consistency). If semantic cues lead to better learning, accuracy for trained nouns should be higher in the fully consistent, followed by the partially consistent, and finally the inconsistent condition. While we included type-frequency as a predictor in the analyses of trained nouns the predictions here are less clear and these analyses should be considered exploratory. If high type-frequency facilitates the identification of the semantic cue, which in turn facilitates recall of cue-consistent particles for trained items, we might expect high type-frequency to facilitate performance in the fully and partially consistent conditions. However, our high type-frequency condition had lower token frequency (i.e., more nouns in each category but fewer exposures to each noun in training), which might lead to *lower* performance on trained nouns in these conditions.

**Children:** Accuracy was higher in Session 4 (74.9%) than Session 1 (56.6%) ( $\beta = 1.12$ ,  $SE = 0.11$ ,  $z = 10.50$ ,  $p < .001$ ), although participants were above chance even in Session 1 ( $\beta = 0.29$ ,  $SE = 0.07$ ,  $z = 4.39$ ,  $p < .001$ ). Performance was significantly higher in

the fully consistent condition than the partially consistent condition ( $\beta = 0.67$ ,  $SE = 0.22$ ,  $z = 3.05$ ,  $p = .002$ ). This contrast interacted with session ( $\beta = 0.90$ ,  $SE = 0.26$ ,  $z = 3.52$ ,  $p < .001$ ), with a significant effect of semantic consistency emerging only in Session 4 (Session 1:  $\beta = 1.65$ ,  $SE = 0.16$ ,  $z = 1.04$ ,  $p = .30$ , fully consistent = 59.1%, partially consistent = 57.0%; Session 4:  $\beta = 1.06$ ,  $SE = 0.31$ ,  $z = 3.46$ ,  $p = .001$ , fully consistent = 84.1%, partially consistent = 71.4%). There was no evidence that performance in the partially consistent condition (64.8%) was higher relative to the inconsistent condition (62.7%) ( $\beta = 0.09$ ,  $SE = 0.21$ ,  $z = 0.43$ ,  $p = .67$ ), indicating no evidence of a benefit from partially consistent semantic cues.

There was no significant effect of type-frequency ( $\beta = -0.29$ ,  $SE = 0.18$ ,  $z = -1.65$ ,  $p = .099$ ), or interaction with either contrast. However, in each case, means are in the direction of stronger performance in the *low* type-frequency condition, where there were more repetitions per noun during training. This was particularly strong where there were no semantic cues available (inconsistent condition) to compensate for the lower number of repetitions under high type-frequency. There was a three-way interaction between the partially consistent/inconsistent contrast, type-frequency and session ( $\beta = -0.97$ ,  $SE = 0.48$ ,  $z = -2.04$ ,  $p = .041$ ). This result is somewhat difficult to interpret: It could reflect that, in the high type-frequency condition, there were fewer exposures to each noun and therefore performance improved less from Session 1 to Session 4 in the inconsistent condition relative to the partially consistent condition where semantic cues can compensate for low token frequency. However, we do not see generalisation of semantics in the partially consistent condition in the novel nouns analyses below, which speaks against this interpretation.

**Adults:** Accuracy was higher in Session 4 (94.1%) compared to Session 1 (81.3%) ( $\beta = 3.12$ ,  $SE = 0.63$ ,  $z = 5.00$ ,  $p < .001$ ), although participants were above chance even in Session 1 ( $\beta = 2.24$ ,  $SE = 0.21$ ,  $z = 10.49$ ,  $p < .001$ ). Accuracy was also higher given low (92.4%) than high type-frequency input (83.4%) ( $\beta = -1.19$ ,  $SE = 0.59$ ,  $z = -2.04$ ,  $p = .042$ ),

presumably due to individual nouns being repeated more often in the smaller languages. Participants were nevertheless above chance even with the larger languages in Session 1 (high type-frequency:  $\beta = 1.47$ ,  $SE = 0.26$ ,  $z = 5.61$ ,  $p < .001$ ).

Performance was significantly higher in the fully consistent condition (94.5%) compared to the partially consistent condition (83.4%) ( $\beta = 2.46$ ,  $SE = 0.76$ ,  $z = 3.23$ ,  $p = .001$ ). As for children, there was no evidence that performance in the partially consistent condition (85.6%) was higher than performance in the inconsistent condition (83.4%) ( $\beta = -0.29$ ,  $SE = 0.67$ ,  $z = -0.44$ ,  $p = .66$ ). There were no significant interactions between consistency, session, and type-frequency.

***Children and Adults Compared:*** We ran a model on the combined adult and child data, with age included as an additional fixed effect.<sup>8</sup> Adults had higher accuracy than children ( $\beta = 2.35$ ,  $SE = 0.24$ ,  $z = 9.92$ ,  $p < .001$ ). Accuracy was higher in Session 4 compared to Session 1 ( $\beta = 1.67$ ,  $SE = 0.16$ ,  $z = 10.48$ ,  $p < .001$ ), and this interacted with age ( $\beta = 1.14$ ,  $SE = 0.31$ ,  $z = 3.67$ ,  $p < .001$ ), with children improving less than adults from Session 1 to 4. As seen in the separate analyses for both children and adults, performance was significantly higher in the fully consistent compared to the partially consistent condition ( $\beta = 1.33$ ,  $SE = 0.27$ ,  $z = 4.85$ ,  $p < .001$ ) and this interacted with age ( $\beta = 1.34$ ,  $SE = 0.57$ ,  $z = 2.37$ ,  $p = .018$ ), with children showing a smaller difference in performance between fully consistent and partially consistent conditions than adults.

While the combined model suggests no difference in performance between partially consistent and inconsistent conditions ( $\beta = -0.05$ ,  $SE = 0.25$ ,  $z = -0.21$ ,  $p = .83$ ), this consistency contrast is involved in a two-way interaction with session ( $\beta = -0.54$ ,  $SE = 0.27$ ,  $z = -2.00$ ,  $p = .046$ ), suggesting greater improvement from sessions 1 to 4 in the inconsistent condition, and a three-way interaction with session and age ( $\beta = -1.18$ ,  $SE = 0.57$ ,  $z = -2.06$ ,  $p = .04$ ), suggesting that this is driven by adults rather than children. Since performance is always numerically higher in the partially consistent condition, we interpret this as (tentative)

evidence that, for adults only, there was an early (i.e. Session 1) benefit for partial consistency which disappears in Session 4, likely due to ceiling effects.

### *Novel Nouns*

Data from novel nouns (Figure 2) were analysed in models predicting particle accuracy by (the fixed factors) *semantic consistency* (fully consistent/partially consistent only – in the inconsistent condition, there was no “correct” or majority particle based on semantic category), *type-frequency* (high/low), *session* (1/4), and their interactions. As for the analysis of trained nouns, we first analyse the data for aware/unaware participants combined. If semantic cues lead to generalisation, we should see above chance performance in both conditions (although we expect accuracy to be highest when cues are fully consistent). We expect greater generalisation given high type-frequency in both conditions.

[Figure 2 about here]

**Children:** Performance improved significantly over time ( $\beta = 0.39$ ,  $SE = 0.09$ ,  $z = 4.15$ ,  $p < .001$ , Session 1 = 52.8%, Session 4 = 60.1%) and more correct particles were produced in the fully consistent (63.2%) than the partially consistent condition (50.7%), ( $\beta = 0.66$ ,  $SE = 0.21$ ,  $z = 3.16$ ,  $p = .002$ ). There was also a significant interaction between session and semantic consistency ( $\beta = 0.50$ ,  $SE = 0.19$ ,  $z = 2.71$ ,  $p = .007$ ), with children producing more correct particles between sessions only in the fully consistent condition ( $\beta = 0.65$ ,  $SE = 0.14$ ,  $z = 4.60$ ,  $p < .001$ , Session 1 = 56.7%, Session 4 = 68.6%); there was no evidence of change between sessions in the partially consistent condition ( $\beta = 0.14$ ,  $SE = 0.12$ ,  $z = 1.15$ ,  $p = .25$ , Session 1 = 49.0%, Session 4 = 52.0%). There was no effect of type-frequency and no interaction with this factor.

Since the comparison with chance is key for novel nouns, we also fitted separate intercepts for each of the conditions in each session (comparing each to 50% chance

performance). Children in the fully consistent condition performed significantly above chance in both sessions (Session 1:  $\beta = 0.35$ ,  $SE = 0.13$ ,  $z = 2.66$ ,  $p = .008$ ; Session 4:  $\beta = 0.99$ ,  $SE = 0.19$ ,  $z = 5.15$ ,  $p < .001$ ), whilst performance in the partially consistent condition did not exceed chance in either session (Session 1:  $\beta = -0.04$ ,  $SE = 0.13$ ,  $z = -0.33$ ,  $p = .75$ ; Session 4:  $\beta = 0.10$ ,  $SE = 0.18$ ,  $z = 0.56$ ,  $p = .58$ ).

**Adults:** Performance improved significantly over time ( $\beta = 2.22$ ,  $SE = 1.02$ ,  $z = 2.17$ ,  $p = .03$ , Session 1 = 78.6%, Session 4 = 87.6%) and participants produced more correct particles in the fully consistent (91.2%) than the partially consistent condition (75.0%) ( $\beta = 3.44$ ,  $SE = 1.03$ ,  $z = 3.34$ ,  $p = .001$ ). The interaction between session and semantic consistency was not significant ( $\beta = 1.89$ ,  $SE = 1.43$ ,  $z = -1.32$ ,  $p = .19$ ) and there was no effect of type-frequency, or interaction with this factor.

Fitting separate intercepts for each condition and each session revealed above chance performance in all cases: fully consistent: Session 1,  $\beta = 4.35$ ,  $SE = 0.83$ ,  $z = 5.22$ ,  $p < .001$ ; Session 4,  $\beta = 7.49$ ,  $SE = 1.46$ ,  $z = 5.15$ ,  $p < .001$ ; partially consistent: Session 1,  $\beta = 1.90$ ,  $SE = 0.66$ ,  $z = 2.89$ ,  $p = .004$ ; Session 4,  $\beta = 3.14$ ,  $SE = 0.93$ ,  $z = 3.38$ ,  $p = .001$ .

**Children and Adults Combined:** As for the trained noun data, we ran a combined model with age as an additional fixed effect, focusing on age-related effects. Children performed worse than adults on novel nouns ( $\beta = 2.73$ ,  $SE = 0.33$ ,  $z = 8.26$ ,  $p < .001$ ). Age interacted with session ( $\beta = 0.95$ ,  $SE = 0.38$ ,  $z = 2.49$ ,  $p = .013$ ) and consistency ( $\beta = 1.77$ ,  $SE = 0.64$ ,  $z = 2.78$ ,  $p = .006$ ), with children improving less than adults from Session 1 to Session 4 and benefitting less from fully consistent input. While the combined data set showed no overall effect of type-frequency, there was a 3-way interaction between type-frequency, consistency and session ( $\beta = 1.35$ ,  $SE = 0.62$ ,  $z = 2.17$ ,  $p = .03$ ), however, this is qualified by a marginal 4-way interaction between type-frequency, consistency, session and age ( $\beta = 2.50$ ,  $SE = 1.34$ ,  $z = 1.87$ ,  $p = .062$ ): while we would be cautious about the reliability of these effects, they are consistent with performance improving more between sessions for the

consistent language at high type-frequency, an effect which is stronger in adults than children. However, as can be seen from Figure 2, rather than reflecting a generalisation advantage for the high type-frequency fully consistent language, this effect shows that adult participants in the low type-frequency fully consistent language are already at ceiling in Session 1 and have little room for improvement over the 4 sessions; thus, we see more improvement in the high type-frequency language.

### ***Summary of Production Data***

Both children and adults exploited fully consistent semantic conditioning cues when learning relationships between nouns and particles. First, they were both able to identify and extend semantic conditioning to novel nouns in the fully consistent condition. Second, fully consistent semantic cues improved both groups' learning performance with trained nouns relative to the two conditions where semantic cues were weaker or absent. This suggests that fully consistent semantic cues can aid learning of trained noun-particle co-occurrences in both children and adults, even though they could have simply relied on co-occurrence when learning these items.

For partially consistent cues, evidence for generalisation with novel nouns was found only in adults. There was also some (rather indirect) evidence from the trained noun analyses where adults (but not children) benefited from partially consistent cues in Session 1. This absence of generalisation in the partially consistent condition for child learners stands in contrast to Samara et al.'s (2017) evidence of (reduced) learning from partially consistent cues. Overall, this suggests that semantic cues may be particularly difficult for children to exploit.

Low type-frequency, where learners encounter each noun more frequently and have a better opportunity to learn individual associations between nouns and particles, benefitted trained noun learning.<sup>9</sup> Most critically however, there was no evidence of the predicted high type-frequency benefit on *generalisation* for either age group.

## 2AFC Test: Accuracy

### *Trained Nouns*

Figure 3 shows the proportion of correct (i.e., attested) particle choices for trained nouns in the 2AFC test. The analyses were identical to those used for production performance (minus the fixed factor Session, since the 2AFC test only took place in Session 4).

[Figure 3 about here]

**Children:** There was a reliable intercept ( $\beta = 1.10$ ,  $SE = 0.13$ ,  $z = 8.33$ ,  $p < .001$ ), indicating that the attested particle was selected with greater than chance (50%) accuracy. Accuracy was significantly higher in the fully consistent (80.0%) compared to the partially consistent condition (68.3%) ( $\beta = 0.73$ ,  $SE = 0.32$ ,  $z = 2.31$ ,  $p = .02$ ). There was no difference in accuracy between the partially consistent and inconsistent conditions, no main effect of type-frequency, and neither consistency contrast interacted with type-frequency.

**Adults:** We simplified the full model due to nonconvergence by removing the interaction between semantic consistency and type-frequency. There was a reliable intercept ( $\beta = 6.09$ ,  $SE = 1.53$ ,  $z = 3.97$ ,  $p < .001$ ) but no main effect of type-frequency and no significant differences between any semantic consistency contrast.

**Children and Adults Combined:** As for the model on adult data, we were required to remove the interaction between consistency and type-frequency due to non-convergence. Children performed worse than adults on trained nouns ( $\beta = 2.25$ ,  $SE = 0.36$ ,  $z = 6.24$ ,  $p < .001$ ), but there were no significant interactions involving age.

### *Novel Nouns*

Figure 4 plots the proportion of semantically appropriate (correct) particle choices for novel nouns. Statistical models were identical to those used for the production data, with the



effect of session (and its interactions) removed. Again, we predicted greater generalisation in the fully consistent condition than in the partially consistent condition.

[Figure 4 about here]

**Children:** Accuracy was higher in the fully consistent (72.1%) compared to the partially consistent condition (50.8%) ( $\beta = 1.25$ ,  $SE = 0.39$ ,  $z = 3.24$ ,  $p = .001$ ). Fitting separate intercepts for each condition revealed that only those in the fully consistent condition were above chance (fully consistent,  $\beta = 1.30$ ,  $SE = 0.29$ ,  $z = 4.53$ ,  $p < .001$ ; partially consistent,  $\beta = 0.05$ ,  $SE = 0.26$ ,  $z = 0.20$ ,  $p = .85$ ). There was no main effect of type-frequency and no interaction between type-frequency and semantic consistency.

**Adults:** We removed the interaction between semantic consistency and type-frequency in the final model due to non-convergence. Accuracy was higher in the fully consistent (97.5%) relative to the partially consistent condition (78.8%) ( $\beta = 3.92$ ,  $SE = 1.39$ ,  $z = 2.81$ ,  $p = .005$ ), however, fitting separate intercepts revealed that performance in both conditions was above chance (fully consistent,  $\beta = 6.53$ ,  $SE = 1.47$ ,  $z = 4.44$ ,  $p < .001$ ; partially consistent,  $\beta = 2.61$ ,  $SE = 0.86$ ,  $z = 3.05$ ,  $p = .002$ ). There was no main effect of type-frequency.

**Children and Adults Combined:** As above, we also ran a combined model with age as an additional fixed effect and simplified by removing the interaction between semantic consistency and type-frequency (due to non-convergence). Children performed worse than adults on novel nouns ( $\beta = 2.81$ ,  $SE = 0.55$ ,  $z = 5.11$ ,  $p < .001$ ). There was a marginal interaction between age and consistency ( $\beta = 1.94$ ,  $SE = 1.05$ ,  $z = 1.85$ ,  $p = .064$ ), with children benefitting less from full consistency. (Note that while the percentage change is similar in adults and children, the change in log-odds space is larger in adults, who are close

to ceiling). There was no overall effect of type-frequency ( $\beta = -0.17$ ,  $SE = 0.43$ ,  $z = -0.39$ ,  $p = .70$ ).

### ***Summary of 2AFC data***

The data from the 2AFC tests mirror the results from the spoken production tests: Children learned the semantic cues only when they were fully consistent, while adults learnt both fully and partially consistent semantic cues. The only difference is that, here, the presence of fully consistent semantic cues influenced children's performance both with trained nouns and novel nouns, while for adults this was true only for novel nouns, however this seems due to ceiling effects across conditions. These findings again indicate learning of fully consistent cues by both age groups, but learning of partially consistent cues by adults only. As in the production test, there was no evidence that type-frequency affected generalisation.

### **Post-Experiment Interview and Explicit Awareness**

[Table 5 about here]

Table 5 shows the number of participants coded as *aware/unaware* in each condition, and awareness is indicated in Figures 1-4 using filled vs. hollow points for individual participants. For both age groups, more participants reported the semantic patterns in the fully consistent than in the partially consistent condition (adults: 19/20 compared with 10/20,  $\chi^2 = 8.03$ ,  $df = 1$ ,  $p = .005$ ; children: 13/30 compared with 2/30,  $\chi^2 = 8.89$ ,  $df = 1$ ,  $p = .003$ ; collapsed across type-frequency).<sup>10</sup> All ten aware adults in the partially consistent condition described both the main categories (animal/vehicle), and the exception items. Only one of the two aware children in the partial condition could report one of the exception items.

Inspection of the individual data in Figures 1-4 suggests that *aware* children and adults show often near-perfect levels of performance. A key question is, thus, whether the group effects were driven by *aware* participants' performance. We test this for novel nouns, i.e., the clearest indicator of participants' ability to use the semantic cue.

### ***Production, novel nouns***

The main analyses revealed significantly above chance performance in the fully consistent condition (both age groups) and the partially consistent condition (adults only). However, Figure 2 suggests that these results may largely be driven by high performance in *aware* participants, with *unaware* participants being clustered around chance. To explore this statistically, where numbers of participants are sufficient<sup>11</sup> (i.e. for children in the fully consistent condition,  $N = 17$ , and adults in the partially consistent condition,  $N = 10$ ) we repeated the statistical analyses on *unaware* participants only: the intercept was not different from chance for either comparison (children, fully consistent:  $\beta = 0.06$ ,  $SE = 0.13$ ,  $z = 0.50$ ,  $p = .62$ ; adults, partially consistent:  $\beta = 0.35$ ,  $SE = 0.20$ ,  $z = 1.74$ ,  $p = .08$ ). There were no significant effects of session or type-frequency, and no session by type-frequency interaction in either model.

### ***2AFC, novel nouns***

Figure 4 plots the 2AFC data for novel nouns; as for production, performance was above chance in the fully consistent condition for both age groups and in the partially consistent condition for adults only. Figure 4 suggests that this again depends on *aware* participants, which was confirmed statistically. Performance did not differ from chance for either *unaware* child participants in the fully consistent condition ( $\beta = 0.10$ ,  $SE = 0.22$ ,  $z = 0.44$ ,  $p = .66$ ), or *unaware* adults in the partially consistent condition ( $\beta = 0.30$ ,  $SE = 0.23$ ,  $z = 1.34$ ,  $p = .18$ ). There was no evidence that performance was modulated by type-frequency in either model.

### *Summary*

Those children and adults who showed evidence of having learned the semantic cues in the production and 2AFC tests (i.e. showed better performance with trained nouns with associated consistent semantic cues and generalised over semantic cues with novel nouns) also verbalized that knowledge in the debrief, indicating explicit awareness. When “aware” participants were excluded from the analyses of the production and 2AFC data, neither adults nor children showed evidence of learning semantic cues.

### **Bayes Factor Analyses**

In the analyses reported above there are several null results which are potentially important. It is, however, difficult to interpret them, since a nonsignificant result ( $p > .05$ ) does *not* tell us whether we have evidence for the null, as opposed to no evidence for any conclusion at all, or even evidence against the null (see Dienes, 2014, for discussion). Therefore, for key null findings, we additionally calculated *Bayes factors* which can be used to assess the strength of evidence for one theory (H1) over another (the null hypothesis).

In each case, we computed the Bayes Factor ( $B$ ) using the method advocated by Dienes (2014; Dienes, Coulton, & Heather, 2018). This requires (i) a model of the data and (ii) a model of H1; (i) comprises an estimate of the effect (i.e. mean difference for the contrast in question) and an estimate of the standard error: we get these from the betas and standard errors of the relevant logistic mixed models, allowing us to meet normality assumptions by continuing to work within log-odds space; For (ii) we model H1 as a half-normal distribution with a mean of 0 and a standard deviation  $x$  which is set to be a rough estimate of the predicted effect (i.e. predicted mean difference for this contrast). This allows for possible effects between 0 and twice the predicted effect, with values closer to 0 being more likely (Dienes, 2014). The notation  $B_{H(0,x)}$  is used (following advice by Dienes: <https://osf.io/hzcv6/>) to denote a  $B$  where the predictions of H1 are modelled as a half normal with an  $SD$  of  $x$ . In the absence of any prior comparable data, and to avoid using unprincipled

default values, we estimated  $x$  for each contrast using values from elsewhere in the data (see Dienes 2019, for similar approach). Details are given for each case below.

We interpret Bayes Factors using the following conventions:  $B < 1/3$  indicates substantial evidence for the null,  $B > 3$  indicates substantial evidence for H1, and values between  $1/3$  and  $3$  indicate that the data do not sensitively distinguish H0 from H1 (Dienes 2008; Jeffreys 1961). Since there is subjectivity in how H1 values are determined, we indicate the robustness of Bayesian conclusions by reporting a robustness region for each  $B$ , which gives the range of values of  $x$  that would qualitatively support the same conclusion (i.e., evidence supporting H0/H1, or, inconclusive evidence). Robustness regions are notated as  $RR [x1, x2]$  where  $x1$  is the smallest  $SD$  that gives the same conclusion and  $x2$  is the largest. They should be interpreted bearing in mind that a larger H0 biases the evidence for the null. Note that for evidence for H0, the maximum  $x$  is always infinity.<sup>12</sup>

### ***No generalisation by children in partial condition?***

For both production and 2AFC performance, the effect of interest is the difference from chance in the partially consistent condition with novel nouns. Therefore, our model of the data in each case is the beta/SE for the intercept in the relevant glmer model. We estimated predicted performance  $x$  to be equal to *half* the equivalent value for children in the consistent condition. This is because performance with consistent cues gives a maximum level of performance we could expect in this condition. Since we model H1  $x$  as the  $SD$  of a half normal distribution with a mean of zero, an  $SD$  of  $x$ , the maximum value is approximately  $2SD$  so  $x$  is half this value. Evidence was ambiguous, both in production ( $\beta = 0.04$ ,  $SE = 0.15$ ,  $B_{H(0,0.35)} = 0.47$ ,  $RR = [0, 0.5]$ ) and 2AFC ( $\beta = 0.05$ ,  $SE = 0.26$ ,  $B_{H(0,0.65)} = 0.43$ ,  $RR = [0, 0.85]$ )).

### ***No benefit of high type-frequency in generalisation?***

We evaluate evidence for the null for higher performance with novel nouns in the high type-frequency than in the low type-frequency condition. We consider only cases where we saw above chance performance overall (i.e., for children the fully consistent condition only, for adults the fully consistent and partially consistent conditions). The effect of interest is the difference between high and low type-frequency conditions and our model of the data, in each case, is the beta/SE for the type-frequency coefficient in the relevant model. In each case, the predicted effect size  $x$  was set to the value of the intercept in the same model, i.e. we are basing our estimate of the main effect on the grand mean (see Dienes, 2019). The logic is as follows: The maximum difference between conditions is seen if low type-frequency participants show chance and high type-frequency participants show performance greater than chance. In this case, if performance on this test is  $p$  (so the grand mean is  $\bar{p}$ ) and chance is  $c$ , the difference in  $p$  between the two conditions will be equal to:  $2(p-c)$ . This gives us an estimate of the maximum value of  $x$ ; since we are using a half normal distribution with a mean of zero, we assume the maximum value is equal to approximately  $2SD$ , so we can set our estimate  $x$  of the standard deviation to be equal to half of this value i.e.  $x = \bar{p} - b$ . Chance here is 50% (i.e. 0 in log-odds space), so we set  $x = \bar{p}$ . For children ( $N = 30$ ), the evidence was ambiguous (production:  $\beta = -0.07$ ,  $SE = 0.44$ ,  $B_{H(0,0.77)} = 0.45$ ,  $RR = [0, 1.07]$ ; 2AFC:  $\beta = 0.62$ ,  $SE = 0.78$ ,  $B_{H(0,1.58)} = 0.87$ ,  $RR = [0, >4.59]$ ). For adults ( $N = 60$ ), there was evidence for the null (production:  $\beta = -0.28$ ,  $SE = 1.01$ ,  $B_{H(0,4.31)} = 0.19$ ,  $RR = [2.32, \text{inf}]$ ; 2AFC:  $\beta = -0.82$ ,  $SE = 1.20$ ,  $B_{H(0,4.57)} = 0.16$ ,  $RR = [2.09, \text{inf}]$ ).

### ***No generalisation by “unaware” children or adults?***

As in previous analyses for awareness, we focus on children in the fully consistent condition and adults in the partially consistent condition. The effect of interest is the difference from chance for *unaware* participants with novel nouns, so our model of the data in each case is the beta/SE for the intercept for the relevant model. We estimate predicted

performance  $x$  to be equal to *half* the equivalent value for *aware* participants. The logic is that performance of *aware* participants gives a maximum level of performance we could expect for *unaware* participants. Since we model  $H_1$   $x$  as the  $SD$  of a half normal distribution with a mean of zero, and  $SD$  of  $x$ , the maximum value is approximately  $2SD$ , so  $x$  is half this value. There was evidence for the null for children (production:  $\beta = 0.06$ ,  $SE = 0.12$ ,  $B_{H(0,0.64)} = 0.29$ ,  $RR = [0.55, \text{inf}]$ ; 2AFC:  $\beta = 0.10$ ,  $SE = 0.22$ ,  $B_{H(0,1.97)} = 0.16$ ,  $RR = [0.95, \text{inf}]$ ). The evidence for adults was ambiguous (production:  $\beta = 0.27$ ,  $SE = 0.16$ ,  $B_{H(0,1.67)} = 0.74$ ,  $RR = [0, 3.75]$ ; 2AFC:  $\beta = 0.30$ ,  $SE = 0.23$ ,  $B_{H(0,2.3)} = 0.43$ ,  $RR = [0, 3]$ ).

### **Summary**

A series of Bayes Factor analyses aimed to clarify the status of three key null results in the frequentist analyses reported above. First, in frequentist analyses, we did not find evidence that children could generalise over partially consistent cues, however, Bayes Factor analyses indicate that the current data are actually ambiguous with respect to this question. Second, the frequentist analyses found no evidence that higher type-frequency benefitted generalisation: the Bayes Factors indicate that the evidence here is ambiguous for children, but adults showed substantial evidence for the null (i.e. evidence that higher type-frequency has *not* benefitted generalisation for these learners). Finally, the frequentist analyses did not find evidence of generalisation in “unaware” learners. Bayes Factor analyses indicate that this was ambiguous for adults, but for children there was evidence for the null. These findings were all consistent across the production and 2AFC tests.

## **General Discussion**

We used a semi-artificial language methodology to explore whether 6-year-olds and adults could learn and generalise relationships between semantic cues associated with nouns and co-occurring grammatical forms. Artificial (or semi-artificial) language methods allow us to isolate semantic cues and determine whether different age groups can exploit them in the

early stages of learning, which is an important step to understand the ability to integrate semantic with other types (e.g., phonological) of cues.

In the current study, both adults and children showed generalisation over fully consistent, i.e., maximally reliable, semantic cues. However, only adults showed learning when the cues were *partially* predictive. Moreover, generalisation of semantic cues to new nouns was not greater in the high type-frequency conditions (where the semantics were exemplified with more instances), for either children or adults, and generalisation appeared to depend on explicit awareness of the cues in question (with one possible exception covered in Appendix C). We discuss these findings and their implications below.

### **Sensitivity to Fully Consistent Semantic Cues**

In our study, *both* children and adults proved able to pick up on fully consistent semantic cues, both with trained and novel test items: They could learn that nouns denoting animals co-occurred with one particle whilst nouns denoting vehicles occurred with another particle. For trained nouns, learning of the noun-particle co-occurrences was boosted compared with matched languages where semantic cues predicted particle choice less reliably (the partially consistent and inconsistent conditions), i.e., it was easier to learn that *cow* was paired with *bup* if all other animal nouns also co-occurred with *bup*. For novel nouns, we saw above-chance usage of the particle consistent with the noun semantics. The advantage of fully consistent input held both for production and 2AFC task performance (except for adults in the 2AFC test where there were no differences for trained nouns due to ceiling effects).

Whilst the finding that adults can learn semantic cues is consistent with previous research (Ferman & Karni, 2010; Ferman et al., 2009; Leung & Williams, 2012; Williams, 2005; Vujovic et al., in press), to our knowledge, only one published study has demonstrated that children can also generalise over isolated fully consistent semantic cues (Culbertson et al., 2019, Experiment 1). In contrast, Ferman and Karni (2010) did not find that young



children (age 8) could generalise across semantic cues, even though they were older than the children we tested and were trained for an additional 11 sessions. What might explain this mismatch in results? First, the pictures that participants saw whilst hearing sentences in the semi-artificial language may have helped them to extract the semantic regularities.

Additionally, although both studies used semi-artificial languages, Ferman and Karni's participants may have focused on information that typically cues gender-class membership in Hebrew. This may have worked against the learning of the novel generalisations. In comparison, English does not have grammatical gender, therefore, our participants did not have to learn a new noun class system that cut across their existing system. Finally, in the current study, we had 30 children for each level of semantic consistency (across the two type-frequency conditions), as opposed to Ferman and Karni's eight participants/age group. Their statistical power is thus low and they do not evaluate the strength of evidence for the null.

Relevant also is that many of our participants do *not* pick up on the semantic cues and that the awareness analysis indicated that group performance was driven by a subset of strong learners; in a study with a much smaller sample size, these participants may simply not occur.

### **Sensitivity to Partially Consistent Semantic Cues**

We also looked at learning of semantics in languages where there were exception items (i.e., all nouns denoting animals occurred with *bup*, except one which occurred with *kem*). Partial consistency is common in natural languages (Mirković et al., 2005). For adults, there was again evidence of generalisation (i.e., above chance usage of the particle which occurred with the majority of nouns with matching semantics) in both the production and 2AFC tests, however, this was significantly weaker than in the fully consistent language.

Children also showed significantly weaker performance in the partially consistent condition relative to the fully consistent condition and, in fact, were not above chance in the generalisation test, although the Bayes factor analysis suggested the data here were

ambiguous (with no evidence for generalisation, but also not for the null, for both 2AFC and production). In fact, looking at Figures 2 and 4, by Session 4, there are two outlier children in the partially consistent condition who *do* show generalisation (and who are also aware of the semantic patterns, a point to which we return below).

Weaker generalisation over partially consistent cues is expected under probabilistic theories of generalisation and is in line with our previous finding of weaker learning of partially consistent cues in a paradigm featuring social (rather than semantic) conditioning, particularly in children (Samara et al., 2017). Recall that Schwab et al. (2018) also manipulated partially reliable semantic cues, though, with an in principle easier manipulation: exception nouns lacked relevant semantic gender cues, rather than cuing the “wrong” gender markers. They also did not find evidence of above chance generalisation in children (although it was not assessed if there was evidence for the null).

Taken together, the findings of these studies suggest that semantic cues are challenging, particularly in the more natural situation in which they are only partially reliable. This is in line with findings from studies of child language acquisition (e.g., Karmiloff-Smith, 1981) and Culbertson et al. (2019, Experiment 2) who demonstrated that, when semantic and phonological cues are equally reliable and are both available from the outset of learning, children prioritize phonological cues.

### **Type-frequency**

We also asked whether the number of nouns that followed a particular pattern would influence the degree to which learners would generalise to novel nouns, and to that end, we compared learning from languages containing more (high type-frequency) versus fewer (low type-frequency) exemplifying nouns. For trained nouns we predicted that item-based learning might be poorer in the larger (high type-frequency) languages, because each noun occurred less often and participants were required to make eight item-by-item associations (rather than

four in the low type-frequency languages). There was partial support for this prediction, but only in adult participants, and only in the production test. In comparison, for novel nouns, we predicted that exposure to more exemplars would lead to greater generalisation with novel nouns (Bybee, 1995; Wonnacott et al., 2012). We found no evidence of stronger generalisation under high type-frequency in any analysis, with one possible exception: as described in Appendix C, in the analysis of trained nouns in the partially consistent condition type-frequency boosted over-generalisation with the *exception nouns* – that is, participants were more likely to use the *wrong* particle with the *exception nouns* (i.e. if the exception noun is *cat*, using the particle that co-occurred with the majority of animals) in the high-type-frequency condition. This could reflect a benefit of high type-frequency for generalisation over semantic cues, i.e., more (over)generalisation when the pattern is exemplified with more nouns. However, there is an alternative explanation in terms of the lower token frequency of the exception items in that condition. See Appendix C for further discussion.

Returning to the main analyses, at least for adults (where  $N = 60$ ), the lack of a main effect of type-frequency is unlikely to be due to power as there was evidence for the null. A more likely explanation is that the paradigm was not picking up on the type of implicit, probabilistic learning and generalisation for which a type-frequency effect is relevant in natural language learning (Apfelbaum & McMurray, 2011; Ramscar et al., 2010), as discussed in the following section.

### **Explicit Awareness**

Post-experiment interviews were used to determine whether participants were aware of the semantic cues present in the input. Many children and adults were able to verbalise the semantic patterns present in the input, and critically, in the novel nouns test, “aware” participants’ performance was very strong (near perfect in some cases), while there was no evidence of above chance generalisation in “unaware” participants (i.e. participants not

reporting any semantic patterns). For child participants, there was evidence for the null for unaware participants, though for adults the evidence was ambiguous.

One possibility – given that the questionnaire was administered after training and testing, is that explicit awareness could be a *product* of successful learning, rather than a determinant of it: Participants may have learned the cues implicitly, then have noticed them explicitly. However, speaking against this, Ferman and colleagues (Ferman & Karni, 2010; Ferman et al., 2009) used a more continuous method for tapping both semantic learning and awareness of semantics cues and did not detect generalisation until around the time participants also showed explicit awareness. Moreover, in our data, the fact that there is no relationship between number of exemplars and generalisation (i.e. no type-frequency effect) is in line with an account where learners are not engaging in implicit, probabilistic learning and generalisation. Ferman et al. (2009) suggest that acquiring the semantic conditioning requires an explicit learning stage that makes use of declarative memory. This is to some extent consistent with Ullman’s dichotomous declarative/procedural neurobiological model of language learning and retention (Ullman 2001, 2016) in which declarative memory subserves learning of arbitrary semantic associations and procedural memory subserves statistical rule learning. In this model, explicit knowledge is always declarative, though declarative knowledge can also be implicit, and there can be interplay and transfer between the two. Declarative memory develops throughout childhood (e.g., Digiulio et al. 1994), potentially explaining the poor learning by children in Ferman and Karni (2010) and in our experiment (though note that in Ferman & Karni, 2010, children also show weaker learning of the phonological form of the rule, which they argue indicates that procedural memory also develops across the age range they study). Further evidence for declarative learning in our data comes from exploratory post hoc analyses presented online (<https://osf.io/sy8zr/>) showing that, for children in the fully consistent condition, verbal working memory (i) is

higher in children classified as “aware” of the semantic cues (ii) predicts performance with novel nouns, at least in the 2AFC task<sup>13</sup> - consistent with the close relationship which has been reported between working memory and declarative learning (Ullman, 2016).

Although our data are consistent with Ullman’s model, it seems very unlikely that there is no implicit learning over semantic cues in natural language learning. Native speakers cannot verbalize knowledge of many word meanings (e.g. function words) and this inability is taken as characteristic of implicit learning (see also discussion in Leung & Williams, 2012). For gender, work by Mirković et al. (2005) has uncovered subtle, probabilistic semantic cues in Serbian (e.g., nouns referring to vegetables tend to be masculine [65%], whilst nouns referring to fruits tend to be feminine [72%]) which play a role in gender classes but are unlikely to be part of the conscious knowledge of a native speaker. It is therefore important to consider why our paradigm does not tap implicit generalisation over semantic cues, especially since other artificial language studies have found evidence of this, at least in adults (Leung & Williams, 2012; Williams, 2005; Vujovic et al., in submission). One possibility is that we did not allow sufficient time for implicit learning, given that formation of procedural memories is predicted to be slower (Ullman 2016). Arguing against this, Ferman and Karni (2010) did not see implicit learning even after 15 training sessions, suggesting it is not simply a question of providing additional exposure. Another possibility is that the nature of our novel noun test specifically draws on explicit learning mechanisms. In support of this, the one place in our data where we *did* see potential evidence of implicit semantic generalisation was in the analyses in Appendix C where we saw *over*-generalisation of semantic cues by unaware participants in the partially consistent condition. It may be easier to tap implicit generalisation in the case where there are “competing” influences on particle choice at test (i.e. the semantic generalisation indicates one particle while the specific noun indicates the other). It is also intriguing that this test was the one place where we did see

a potential type-frequency effect, as expected under probabilistic generalisation (but see Appendix C, for an alternative interpretation). It may also be that offline test measures are more appropriate for eliciting responses based on implicit knowledge – for example, the RT measure used in Leung and Williams, (2012) and Williams (2005) (with adult participants). Finally, it might be that changes need to be made to the training paradigm itself: Vujovic et al. (in submission) (again with adults) used 2AFC tests similar to ours, but the artificial language referred to novel objects with novel semantic features and they employed a speeded presentation during training (1000ms between trials) in contrast to our own self-paced presentation method. These differences may work to discourage explicit strategies and boost implicit learning. Future artificial language learning work should establish child-friendly paradigms which encourage implicit processes.

### **The influence of prior knowledge**

There are several ways that prior experience with natural languages might have influenced our results. First, all of our adult participants had experience with at least one language with grammatical gender (e.g. French or German, encountered during their high school education), whereas children were largely monolingual English speakers. This confound between age and experience with a language with grammatical gender is extremely difficult to avoid in the UK and have might contributed to the fact that adults outperformed children. However, better learning in adults is generally not unexpected in artificial language learning research even in cases where prior knowledge might be expected to have less relevance: for instance, in Samara et al. (2017), adults outperformed children in learning conditioning on *talker* for plurals, which adults would not have encountered in their natural language input.

It is also worth noting that, in the languages our adult participants will typically have encountered, the main semantic cue is natural gender, not animacy. Thus, any advantage must

stem from a more general expectation that languages have gender systems or that grammatical gender can depend on semantics, rather than specific prior knowledge of the semantic cue we used here. On the other hand, prior knowledge could also potentially lead to *poorer* learning. For example, we have discussed how in Ferman and Karni's (2010) study, participants prior knowledge of gender in Hebrew might have *hindered* learning of gender categories based on novel (different) cues.

A second possible concern is that participants' prior experience with linguistic or perceptual categories in general (i.e. not pertaining to grammatical gender) may influence learning. For example, our partially consistent language might be challenging because the gender-based grouping cuts across pre-existing perceptual and linguistic categories (animals and vehicles). To counteract this concern, one could use a fully artificial language, with novel lexical items and referents, though it is unclear whether a fully artificial version of our experiment would produce different results for partially consistent cues. First, Schwab et al. (2018) sidestep this problem by having exception items drawn from a third category and still find that children fail to learn the semantic cue. Second, Culbertson et al. (2019) use a fully artificial lexicon and less familiar referents (novel cartoon planets and aliens) and show results highly similar to ours (i.e., modest above-chance generalisation on novel nouns after 1 session by children trained on a language with fully consistent cues). It is also again possible that using familiar categories could have the reverse effect of *increasing* difficulty: learned knowledge of the animal/vehicle distinction and how it maps onto familiar linguistic forms (e.g. the words "animal" and "vehicle") might block (Kamin, 1968) implicit learning of the mappings to *novel* linguistic forms. In line with this, Vujovic et al. (in submission) did find implicit learning over semantics in adults trained on a fully artificial language with fully novel referents, although there were also other potentially important methodological differences, e.g. speeded training.

A final note of caution for future work with fully artificial stimuli is that mastering new lexical items increases overall memory burden during the experiment. Our own pilot work and several published studies show that in such circumstances learners – especially children – are less likely to produce variable output (i.e. they only produce a single particle or over-produce one particle: Hudson Kam & Chang, 2009; Ferdinand, Kirby & Smith, 2018). Participants who cannot produce variable output cannot provide evidence that they are successfully learning the conditioning cue, thus, using a semi-artificial language and familiar referent categories may be more appropriate.

### **Conclusion**

We showed that both child and adult learners are able generalise words to novel linguistic contexts based on fully consistent semantic cues, though performance was weaker in children. In the more naturalistic situation where the cues are only partially consistent, generalisation was weaker for both groups, and for children there was no evidence that performance was above chance. In both age groups, successful generalisation with novel nouns was accompanied by an ability to verbalize the relationship between the semantic cue and particle choice at the end of the experiment. There was also no clear evidence that generalisation was affected by the number of exemplifying nouns in the input, as would be expected under implicit probabilistic generalisation.

The finding that children have greater difficulty with semantic cues than adults coincides with data from natural language learning suggesting that children are slow to learn semantic cues to noun class. While this is in line with an account in which explicit declarative memory is central in semantic learning, we note that native speakers do acquire implicit knowledge of semantic cues. Thus, future work should both use tasks which promote implicit learning and also directly compare children's learning of semantic and other (e.g., phonological) cues.



### **Acknowledgements**

This research was supported by an Economic and Social Research Council Grant (ES/K006339) awarded to KS and EW. We thank Tarini Mehrotra for assistance with data collection and input.

### References

- Alloway, T. P. (2008). *Automated Working Memory Assessment*. London: Pearson Assessment.
- Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*, 1105-1138. <https://doi.org/10.1111/j.1551-6709.2011.01181.x>
- Arias-Trejo, N., & Alva, E. A. (2013). Early Spanish grammatical bootstrapping: learning nouns through adjectives. *Developmental Psychology*, *49*, 1308-1314. <https://doi.org/10.1037/a0029621>
- Aurilio, M. K., Jenkins, J. J., & Silliman, E. R. (2000). Partial phonological marking facilitates the acquisition of noun subclasses: A replication. *Psychological Reports*, *86*, 975-994. <https://doi.org/10.2466/pr0.2000.86.3.975>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Barbaud, P., Ducharme, C., & Valois, D. (1982). D'un usage particulier du genre en canadien-français: la féminisation des noms à initiale vocalique. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, *27*, 103-133. <https://doi.org/10.1017/S0008413100023860>
- 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*, *44*, 255-278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The Crosslinguistic Study of Sentence Processing* (pp. 3-76) Cambridge University Press: New York, USA.

- Bates, D., Maechler, M., & Bolker, B. (2013). lme4: Linear mixed-effects models using Eigen and Eigenfaces. R package version 0.999999-0.
- Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.
- Braine, M.D. S., Brody, R. E., Brooks, P. J., Sudhalter, V., Ross, J. A., Catalano, L., & Fisch, S. M. (1990). Exploring language acquisition in children with miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language*, 29, 591-610. [https://doi.org/10.1016/0749-596X\(90\)90054-4](https://doi.org/10.1016/0749-596X(90)90054-4)
- Braine, M. D. S. (1987). What is learned in acquiring word classes: A step towards an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 65-88). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, 32, 76-95. <https://doi.org/10.1006/jmla.1993.1005>
- Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10, 425-455. <https://doi.org/10.1080/01690969508407111>
- Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, 63, 121-170. [https://doi.org/10.1016/S0010-0277\(96\)00793-7](https://doi.org/10.1016/S0010-0277(96)00793-7)
- Corbett, G. G. (1991). *Gender*. Cambridge, England: Cambridge University Press. <https://doi.org/10.1017/CBO9781139166119>
- Culbertson, J., Gagliardi, A., & Smith, K. (2017). Competition between phonological and semantic cues in noun class learning. *Journal of Memory and Language*, 92, 343-358. <https://doi.org/10.1016/j.jml.2016.08.001>

- Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2019). Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias. *Language*, 95, 268-293. <https://doi.org/10.1353/lan.0.0234>
- Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Basingstoke: Palgrave Macmillan.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, 781. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dienes, Z., Coulton, S., and Heather, N. (2018) *Using Bayes Factors to evaluate evidence for no effect: examples from the SIPS project*. *Addiction*, 113 (2). pp. 240-246. <http://dx.doi.org/10.1111/add.14002>
- Digiulio, D. V., Seidenberg, M., O'Leary, D. S., & Raz, N. (1994). Procedural and declarative memory: A developmental study. *Brain and Cognition*, 25, 79-91. <https://doi.org/10.1006/brcg.1994.1024>
- Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, 184, 53-68. <https://doi.org/10.1016/j.cognition.2018.12.002>
- Ferman, S., & Karni, A. (2010). No childhood advantage in the acquisition of skill in using an artificial language rule. *PLoS ONE*, 5, e13648. <https://doi.org/10.1371/journal.pone.0013648>
- Ferman, S., Olshtain, R., Schechtman, E., & Karni, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics*, 22, 384-412. <https://doi.org/10.1016/j.jneuroling.2008.12.002>
- Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, 39, 218-245. <https://doi.org/10.1006/jmla.1998.2569>

- Gagliardi, A., Feldman, N. H., & Lidz, J. (2017). Modeling statistical insensitivity: Sources of suboptimal behaviour. *Cognitive Science*, *41*, 188-217.  
<https://doi.org/10.1111/cogs.12373>
- Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, *90*, 58-89. <https://doi.org/10.1353/lan.2014.0013>
- Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249-268.  
<https://doi.org/10.1017/S0305000904006786>
- Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-436. <https://doi.org/10.1111/1467-9280.00476>
- Grimshaw, J. (1981). Form, function, and the language acquisition device. In C.L. Baker & J. McCarthy (Eds.), *The Logical Problem of Language Acquisition* (pp.165-178). MIT Press: Cambridge, MA, USA.
- Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 815-821.  
<https://doi.org/10.1037/a0015097>
- Hudson Kam, C.L. & Newport, E.L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151-195. <https://doi.org/10.1080/15475441.2005.9684215>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434-446.  
<https://doi.org/10.1016/j.jml.2007.11.007>
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

- Kamin, L. J. (1968). Attention-like processes in classical conditioning. In M. R. Jones, (Ed.), *Miami Symposium. Predictability, Behavior and Aversive Stimulation*. Miami: University of Miami Press.
- Karmiloff-Smith, A. (1981). *A functional approach to child language: A study of determiners and reference* (Volume 24). New York, NY: Cambridge University Press.
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavioural Research Methods*, 44, 978-990.  
<https://doi.org/10.3758/s13428-012-0210-4>
- Lany, J., & Saffran, J. R. (2010). From statistics to meaning: infants' acquisition of lexical categories. *Psychological Science*, 21, 284-291.  
<https://doi.org/10.1177/0956797609358570>
- Lany, J., & Saffran, J. R., (2011). Interactions between statistical and semantic information in infant language development. *Developmental Science*, 14, 1207-1219.  
<https://doi.org/10.1111/j.1467-7687.2011.01073.x>
- Leung, J. H. C., & Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning*, 62, 634-662.  
<https://doi.org/10.1111/j.1467-9922.2011.00637.x>
- Li, P. & MacWhinney, B. (2013). Competition model. In C.A. Chapelle (Ed.). *The Encyclopaedia of Applied Linguistics*. Blackwell Publishing Ltd.  
<https://doi.org/10.1002/9781405198431.wbeal0168>
- Maratsos, M. (1982). The child's construction of grammatical categories. In E. Wanner & L. R. Gleitman (Eds.), *Language Acquisition: The State of the Art* (pp.240-266). Cambridge: Cambridge University Press.

- Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26, 393-424. [https://doi.org/10.1207/s15516709cog2604\\_1](https://doi.org/10.1207/s15516709cog2604_1)
- Mintz, T. H., Wang, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (bigram) parts. *Cognitive Psychology*, 75, 1-27. <https://doi.org/10.1016/j.cogpsych.2014.07.003>
- Mirković, J., Forrest, S., & Gaskell, M. G. (2011). Semantic regularities in grammatical categories: Learning grammatical gender in an artificial language. In L. Carlson, C. Holscher, & T. Shiple (Eds). *Proceedings of the 33<sup>rd</sup> Annual Conference of the Cognitive Science Society* (pp.324-329). Austin, TX: Cognitive Science Society.
- Mirković, J., MacDonald, M. C., & Seidenberg, M. S. (2005). Where does gender come from? Evidence from a complex inflectional system. *Language and Cognitive Processes*, 20, 139-167. <https://doi.org/10.1080/01690960444000205>
- Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, 96(2), 143-182. <https://doi.org/10.1016/j.cognition.2004.09.001>
- Morgan, J. L., & Demuth, K. (2014). *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Psychology Press. <https://doi.org/10.4324/9781315806822>
- Mulford, R. (1985). Comprehension of Icelandic pronoun gender: Semantic versus formal factors. *Journal of Child Language*, 12, 443-453. <https://doi.org/10.1017/S030500090000653X>
- Pérez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. *Journal of Child Language*, 18, 571-590. <https://doi.org/10.1017/S0305000900011259>

- Perfors, A., Tenenbaum, J., & Wonnacott, E. (2009). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*, 607-642.  
<https://doi.org/10.1017/S0305000910000012>
- Pinker, S. (1984). *Language Learnability and Language Development* (1984). Cambridge, MA: Harvard University Press. <https://doi.org/10.1017/S0272263100007117>
- Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 43-102. [https://doi.org/10.1016/0010-0277\(91\)90022-V](https://doi.org/10.1016/0010-0277(91)90022-V)
- Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children with connectionist nets. *Cognition*, *48*, 21-69.  
[https://doi.org/10.1016/0010-0277\(93\)90057-3](https://doi.org/10.1016/0010-0277(93)90057-3)
- Powell, M. J. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives*. Cambridge.
- Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modelling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413-425. <https://doi.org/10.1016/j.jml.2008.02.002>
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). Feature-label-order effects and their implications for symbolic learning. *Cognitive Science*, *34*, 909-957.  
<https://doi.org/10.1111/j.1551-6709.2009.01092.x>
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425-469.  
[https://doi.org/10.1207/s15516709cog2204\\_2](https://doi.org/10.1207/s15516709cog2204_2)
- Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). Distributional learning of subcategories in an artificial grammar: Category generalisation and subcategory



restrictions. *Journal of memory and language*, 97, 17-29.

<https://doi.org/10.1016/j.jml.2017.07.006>

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44(4), 493-515. <https://doi.org/10.1006/jmla.2000.2759>

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially-conditioned linguistic variation. *Cognitive Psychology*, 94, 85-114.

<https://doi.org/10.1016/j.cogpsych.2017.02.004>

Schwab, J. F., Lew-Williams, C., & Goldberg, A. E. (2018). When regularization gets it wrong: Children over-simplify language input only in production. *Journal of Child Language*, 45, 1054-1072. <https://doi.org/10.1017/S0305000918000041>

Smith, K. H. (1969). Learning co-occurrence restrictions: Rule induction or rote learning? *Journal of Verbal Learning and Verbal Behavior*, 8, 319-321.

[https://doi.org/10.1016/S0022-5371\(69\)80086-1](https://doi.org/10.1016/S0022-5371(69)80086-1)

Szagan, G., Stumper, B., Sondag, N., & Franik, M. (2007). The acquisition of gender marking by young German-speaking children: Evidence for learning guided by phonological regularities. *Journal of Child Language*, 34, 445-471.

<https://doi.org/10.1017/S0305000906007951>

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature reviews neuroscience*, 2(10), 717-726.

<https://doi.org/10.1038/35094573>

Ullman, M. T. (2016). The declarative/procedural model: a neurobiological model of language learning, knowledge, and use. In *Neurobiology of language* (pp. 953-968). Academic Press. <https://doi.org/10.1016/B978-0-12-407794-2.00076-6>

- Vigliocco, G., Vinson, D. P., Martin, R. C., & Garrett, M. F. (1999). Is "count" and "mass" information available when the noun is not? An investigation of tip of the tongue states and anomia. *Journal of Memory and Language*, 40, 534-558.  
<https://doi.org/10.1006/jmla.1998.2626>
- Vujović, M., Ramscar, M., & Wonnacott, E. (under review). Language learning as uncertainty reduction: The role of prediction error in linguistic generalisation and item-learning. Pre-print at <https://osf.io/f2n9d/>
- Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, 27, 269-304. <https://doi.org/10.1017/s0272263105050138>
- Wonnacott, E. (2011). Balancing generalisation and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, 65(1), 1-14.  
<https://doi.org/10.1016/j.jml.2011.03.001>
- Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, 66, 458-478. <https://doi.org/10.1016/j.jml.2011.11.004>
- Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language*, 95, 36-48.  
<https://doi.org/10.1016/j.jml.2017.01.005>

## Footnotes

<sup>1</sup> Assignment to condition is random so any differences in memory performance between conditions is accidental. We explore the data for accidental differences in the supplementary materials at <https://osf.io/sy8zr/>. Versions of each of the analyses for adults and children (for trained and untrained nouns in Production and AFC tests) with each of the three memory measures as an additional predictor (12 models in total) are also included. Importantly, the qualitative pattern of results did *not* change in any case, and there was only one case where a significant result became non-significant (in the model for children, trained nouns, AFC task- the contrast between the fully consistent and partially consistent conditions).

<sup>2</sup> All adult participants reported having learned at least one language with gender classes at school (in the UK foreign language teaching was compulsory from age 11 up to 2014; the age has subsequently been lowered to 7; the majority of schools teach romance languages (French, Spanish) or German - [researchbriefings.files.parliament.uk/documents/CBP-7388/CBP-7388.pdf](https://researchbriefings.files.parliament.uk/documents/CBP-7388/CBP-7388.pdf)). None of the children reported any knowledge of a language with grammatical gender beyond a few words.

<sup>3</sup> The full set of stimuli was: Animals – *cow, dog, elephant, fox, giraffe, hamster, hedgehog, hippo, kangaroo, panda, pig, rabbit, sheep, squirrel, tiger, zebra*; Vehicles – *ambulance, bike, boat, bus, car, digger, fire-engine, helicopter, plane, rocket, scooter, tank, tractor, train, truck, van*; Particles – *bup, dak, fod, gos, jeb, kem, pag, tid, wib, yav*).

<sup>4</sup> Using two referents provides a *prima facie* function for the particle, namely that it marks duality/plurality, and allowed us to present nouns in the absence of the particle during the noun practice phases by presenting referents singly.

<sup>5</sup> During test trials, synonymous labels (e.g., using *lorry* instead of *truck*) were accepted as correct to minimise data loss.

<sup>6</sup> We tested only eight trained nouns since this is the total number of items included during training in the low type-frequency condition. In the high type-frequency condition four animals and four vehicles were randomly selected from the set of 16 trained items.

<sup>7</sup> These trials were double coded (by the first and third authors) and where possible were re-coded to match one of the trained particles. Inter-rater agreement on these items was high (Cohen's Kappa = 0.99).

<sup>8</sup> In addition to the age-related effects we focus on in the main text, the combined model indicates a significant effect of type-frequency ( $\beta = 0.66$ ,  $SE = 0.21$ ,  $z = 3.07$ ,  $p = .002$ ), with higher accuracy on the low type-frequency language. This is presumably driven by the higher token frequency in the smaller languages. In the separate analyses of child and adult data this type-frequency effect shows up for adults but not children; this is reflected in the combined model in a (highly marginal) interaction between type-frequency and age ( $\beta = -0.76$ ,  $SE = 0.45$ ,  $z = -1.69$ ,  $p = .091$ ).

<sup>9</sup> As described in Appendix C, in the analysis of trained nouns in the partially consistent condition, we saw an effect of *type-frequency* in interaction with *noun-type*, i.e. whether the noun was one that occurred with the majority-particle for that semantic type (e.g. in a language where most nouns denoting animals co-occur with *bup*, *cat* co-occurring with *bup*) or was an exception (e.g. *dog* co-occurring with *kem*). Specifically, there was more overgeneralisation with the exception nouns in the high type-frequency condition. However, it is unclear whether this reflects greater generalisation in the high type-frequency condition or is driven by the lower token frequency of the exception items in that condition. See Appendix C for discussion.

<sup>10</sup> An equivalent analysis of the effects of type-frequency on awareness (collapsing across consistency) indicates no significant differences between high and low type-frequency (adults: 14/20 aware at high type-frequency compared with 15/20 at low type-frequency,  $\chi^2 = 0$ ,  $df = 1$ ,  $p = 1$ ; children: 10/30 aware at high type-frequency compared with 5/30 at low type-frequency,  $\chi^2 = 1.42$ ,  $df = 1$ ,  $p = .233$ ).

<sup>11</sup> Note that there was only one unaware adult in the fully consistent condition, preventing us from running statistical tests (although this participant performed close to the chance proportion); similarly, there were only two aware children in the partially consistent condition (but these were the two best performers on session 4).

<sup>12</sup> To find the robustness regions, we tested values of  $x$  which are reasonable given the scale, specifically 100 steps from 0 in log odds to 4.595 in log odds space (corresponding to odds/odds ratio of 1.041 which would be the case if comparing two groups where one group was at chance and the other had near perfect performance at 99% accuracy).

<sup>13</sup> However there are also relationships with non-verbal short-term memory which are not straightforwardly accounted for in this mod

**Appendix A: Summary of previous research, including age group(s) tested, type of cue (semantic, phonological, and/or distributional), cue consistency (fully consistent, partially consistent, and/or inconsistent), stimuli used (natural, artificial, or semi-artificial), and a short summary of the key finding(s).**

Author(s)	Date	Age Group(s)	Type of Cue	Cue Consistency	Language	Stimuli	Main Findings
Aurilio, Jenkins, & Silliman	2000	Adults	Phonological	Partially consistent; Inconsistent	Artificial	Name + Noun + Locative Suffix (e.g. <i>Frippy roik-eff</i> ); Different locative suffixes were used for each noun subclass.	Evidence of noun subclass learning and generalisation based on partially consistent phonological cues
Braine	1987	Adults	Semantic (natural gender)	Partially consistent; Inconsistent	Artificial	Noun + Number word	Evidence of noun subclass learning and generalisation based on partially consistent semantic cues
Braine, Brody, Brooks, Sudhalter, Ross, Catalano, & Fisch	1990	7-10 year olds	None	NA	Artificial	Name + Noun + Locative Suffix (e.g. <i>Frippy wern-tev</i> )	Children did not learn to distinguish noun subclasses when these subclasses were arbitrary
Brooks, Braine, Catalano, Brody, & Sudhalter	1993	9-11 year olds; adults	Phonological	Partially consistent; Inconsistent	Artificial	Name + Noun + Locative Suffix (e.g. <i>Frippy choik-eff</i> ); Different locative suffixes were used for each noun subclass.	Evidence of noun subclass learning and generalisation based on partially consistent phonological cues in both age groups; Adult participants also demonstrated explicit knowledge of the word classes.
Culbertson, Gagliardi, & Smith	2017	Adults	Semantic (animacy, shape, flexibility); Phonological	Fully consistent	Artificial	Prefix + Noun + Suffix + Marker (e.g. <i>di-vok-te kuh</i> )	Participants relied more on early-learned cues, even when later learned cues were more salient.
Culbertson, Jarvinen, Haggarty, & Smith	2019	6-7 year olds; adults	Semantic (animacy); Phonological	Fully consistent	Artificial	Noun + Plural Marker (which could be cued by semantics, phonology, or both, e.g. <i>mata kuh</i> )	Evidence of noun subclass learning and generalisation in both children and adults when cues were presented in isolation

							and were fully consistent. However, when cues were available simultaneously adults prioritised semantic cues whilst children prioritised phonological cues.
Ferman & Karni	2010	8 year olds; 12 year olds; Adults	Semantic (animacy)	Fully consistent	Semi-artificial	Hebrew noun + Hebrew verb + novel suffix indicating animacy of noun (e.g. <i>harofe tip-év</i> )	Only 12-year olds and adults were able to generalise the artificial morphological rule to novel items; 7/8 participants in each of these age groups could verbalise the rule explicitly.
Ferman, Olshain, Schechtman, & Karni	2009	Adults	Semantic (animacy)	Fully consistent	Semi-artificial	Hebrew noun + Hebrew verb + novel suffix indicating animacy of noun (e.g. <i>harofe tip-év</i> )	Participants generalised the artificial morphological rule to novel items. Explicit awareness of the semantic aspect of the rule coincided with an abrupt increase in accuracy on generalisation items.
Frigo & McDonald	1998	Adults	Phonological	Partially consistent; Inconsistent	Artificial	Time of day [Morning/Evening] + Noun + Marker + Greeting (e.g., <i>Morning, opr-ash jai</i> )	Evidence of noun subclass learning and generalisation based on partially consistent phonological cues
Gagliardi & Lidz	2014	4-7 year olds; 8-12 year olds; Adults	Semantic (biological gender, physical properties of the noun); Phonological	Natural Language	Natural (Tsez)	Noun-verb agreement elicited production task. Real words and non-words marked with semantic and/ or phonological features	Evidence that all age groups use semantic and phonological cues to classify both real words and non-words. When cues are in conflict, children are more likely to use phonological cues.
Gerken, Wilson, & Lewis	2005	17 month olds	Phonological	Partially consistent	Natural (Russian)	Gender categories: Noun (with phonological marker at end) + Case Marker (e.g., <i>stroj-tel'ya</i> )	Infants could discriminate grammatical from ungrammatical items, but only when a subset of the training items were double-marked with phonological cues to gender category.

Karmiloff-Smith	1981	3-12 year olds	Semantic (natural gender); Phonological	Natural Language	Natural (French)	Gender categories: <i>Exp7</i> – article gender and noun suffix matched; <i>Exp8</i> – article gender not provided; <i>Exp9</i> – article gender and noun suffix mismatched; <i>Exp10</i> – gender of person and noun suffix mismatched (no article); <i>Exp11</i> – given name of male picture, asked to generate female version	Until age 10, children use phonological rather than semantic cues to determine gender classes in their first language
Lany & Saffran	2010	22 month olds	Semantic (animacy); Phonological; Distributional	Fully consistent; Inconsistent	Artificial	Phonological Marker + Noun (e.g. <i>ong coomo</i> ); noun classes were also marked by length (monosyllabic vs. bisyllabic); one noun subclass was paired with pictures of animals, the other with vehicles.	Infants learned and generalised the associations between semantic and phonological/distributional cues and the noun subclasses
Lany & Saffran	2011	22 month olds	Semantic (animacy); Phonological; Distributional	Fully consistent; Inconsistent	Artificial	Phonological Marker + Noun (e.g. <i>ong coomo</i> ); noun classes were also marked by length (monosyllabic vs. bisyllabic); one noun subclass was paired with pictures of animals, the other with vehicles.	Generalisation of noun subclasses depended on an infant's language proficiency; Smaller vocabulary – generalisation based on phonological cues; Larger vocabulary – generalised based on distributional cues
Leung & Williams	2012	Adults	Semantic (animacy)	Fully consistent	Semi-artificial	Novel determiner + English noun (e.g., <i>ro bull</i> ). Different determiners were used for animate and inanimate nouns ( <i>Exp1</i> ) or different sizes ( <i>Exp2</i> )	Unaware participants responded slower in an animacy decision tasks when the mapping between the determiner and noun was incorrect, indicating implicit learning of the noun subclasses. This was not the case for decisions about relative size.
Mintz, Wang, & Li	2014	Adults	Distributional	NA	Artificial	Three-word sequences in which first and last word formed a frame (e.g., <i>choon lartsu blit</i> )	Evidence of learning from distributional cues (“frequent frames”) in the absence of other

							cues (e.g. phonological / semantic)
Mirković, Forrest, & Gaskell	2011	Adults	Semantic (natural gender); Phonological	Fully consistent	Artificial	Determiner + noun + suffix (e.g. <i>tib scoiff-esh</i> )	Participants selected the correct determiner for novel items more often when the semantic (i.e. gender) and phonological (i.e. suffix) cues matched.
Mulford	1985	4-8 year olds	Semantic (natural gender); Phonological	Natural Language	Natural (Icelandic)	Pronoun categorisation test: (e.g. <i>Hvar er hún? – Where is she?</i> ); included real words and non-words	Young children showed greater sensitivity to semantic than phonological cues.
Perez-Pereira	1991	4-11 year olds	Semantic (natural gender); Morphophonological; Syntactic	NA	Artificial	Noun + colour word (which is marked for gender, e.g. <i>rojo/roja – red</i> )	The children paid more attention to syntactic and morphophonological information than to semantic information
Schwab, Lew-Williams, & Goldberg	2018	6 year olds; Adults	Semantic (natural gender)	Partially consistent	Semi-artificial	Moop + novel classifier _ English noun (e.g. <i>moop dax boy; moop po girl</i> )	Adults generalised the semantic cue to untrained items, children did not.
Williams	2005	Adults	Semantic (animacy)	Fully consistent	Semi-artificial	Determiner + English noun (e.g. <i>gi dog, ro book</i> )	Evidence of generalisation based on semantic cues in both aware and unaware participants.

**Appendix B: Questions and Prompts used in the Post-Experiment Interview**

(1) Do you speak any other languages? Do you think that knowing these languages helped you to learn how to speak like Freddy Frog?

(2) What do you think “glim” means?

(3) What do you think X and Y mean (where X and Y are the two particles)?

(4) How did you decide when to use X and when to use Y?

(5) Did you notice any patterns in how Freddy Frog used X and Y? How early in the experiment did you notice these patterns?

(6) Adults Only: Did you consider any other possible patterns along the way that turned out to be incorrect?

(7) Adults Only: Did you focus more on learning which words went with each individual items (e.g., which word went with dog), or on which words went with which *type* of item (e.g., which word went with animals)?



### **Appendix C: Additional analyses looking at exception versus majority items in the partially consistent conditions**

This appendix contains a series of analyses focussing on data from the partially consistent condition. Specifically, we compare performance on exception versus majority-particle nouns (i.e. where all but one of the vehicle nouns occur with particle 1, vehicle nouns occurring with particle 1 are majority items and the vehicle noun occurring with particle 2 is the exception noun) in trained noun tests. For exception items a response is scored as correct if the participant produced the category-exceptional particle, indicating that they had learned the exception. A benefit for the majority-particle nouns could indicate that semantic cues affected performance in the partially consistent condition: learners who are sensitive to the relevant semantic cue should perform better on nouns which conform to the semantic cue predicting particle choice.

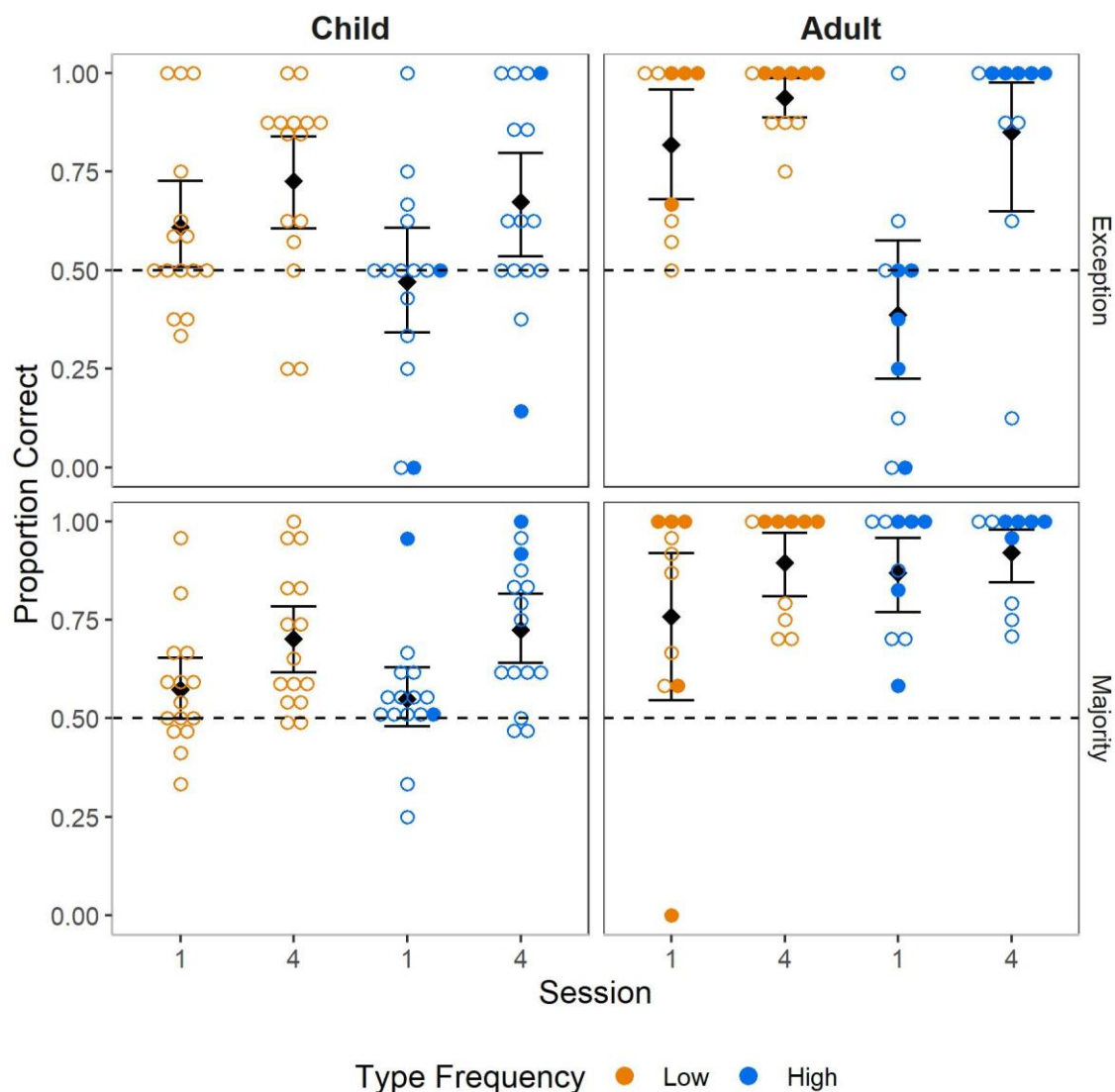
For these analyses we ran models containing the fixed factors *noun-type* (exception noun/majority-particle noun), *type-frequency* (low/high), *session* (1/4), and all interactions between these factors, with random effects and factor coding as described for the analyses in the main text (positive beta indicates: higher performance in Session 4; higher performance in high type-frequency; higher performance with majority nouns).

#### **Production Data**

**Children:** The full model failed to converge – we found a simplified converging model by removing correlations between slopes. As in the trained nouns analysis in the main text, there was a significant effect of session ( $\beta = 0.74$ ,  $SE = 0.16$ ,  $z = 4.62$ ,  $p < .001$ ) but no significant effect of type-frequency or session by type-frequency interaction. Type-frequency didn't interact with any other factor. There was no significant difference between exception and majority-particle nouns ( $\beta = 0.00$ ,  $SE = 0.18$ ,  $z = 0.00$ ,  $p = .998$ ), and no interactions with session, type-frequency, or session by type-frequency. This pattern of results is indicative of a

lack of sensitivity to the semantic cue in the partially consistent condition, which aligns with the results of the between-condition comparisons in the main text.

**Adults:** The full model failed to converge – we found a simplified converging model by removing correlations between slopes. As in the analysis in the main text, there was a significant effect of session ( $\beta = 1.53$ ,  $SE = 0.58$ ,  $z = 2.63$ ,  $p = 0.008$ ); unlike in the main analyses the benefit for the low type-frequency language was not significant. Again, the session by type-frequency interaction was *ns*. Noun-type was significant ( $\beta = 0.85$ ,  $SE = 0.26$ ,  $z = 3.21$ ,  $p = .001$ ), indicating higher performance with the majority than the exception nouns, however this was qualified by two interactions: First, a significant noun-type by session interaction ( $\beta = -1.70$ ,  $SE = 0.85$ ,  $z = -2.01$ ,  $p = .044$ ), reflecting significantly greater performance with the majority-noun than the exception noun in Session 1 ( $\beta = 1.74$ ,  $SE = 0.48$ ,  $z = 3.64$ ,  $p < .001$ ) but not in Session 4 ( $\beta = 0.04$ ,  $SE = 0.51$ ,  $z = 0.86$ ,  $p = .931$ ), likely due to ceiling effects. Second, a noun-type by type-frequency interaction ( $\beta = 2.29$ ,  $SE = 0.53$ ,  $z = 4.33$ ,  $p < .001$ ), reflecting greater performance with the majority noun in the high type-frequency condition ( $\beta = 1.96$ ,  $SE = 0.37$ ,  $z = 5.29$ ,  $p < .001$ ), but not the low type-frequency condition ( $\beta = -0.33$ ,  $SE = 0.38$ ,  $z = -0.88$ ,  $p = .38$ ) i.e., weaker learning of the exception item in the language with a greater number of nouns exemplifying the majority pattern. Inspecting figure C1, it looks as though the reduced learning of the exception nouns is in fact restricted to Session 1 for the high-type-frequency condition, however the 3-way interaction between noun-type, type-frequency, and session was not significant ( $\beta = -2.39$ ,  $SE = 1.73$ ,  $z = -1.38$ ,  $p = .17$ ).



*Figure C1:* Proportion of correctly produced particles for trained nouns, partially-consistent condition only. Diamonds and error bars indicate means and bootstrapped 95% confidence intervals. Each coloured point represents a single participant: as in Figure 1, filled points represent aware participants, unfilled points represent unaware participants. Dashed lines show chance performance. Performance on majority-particle and exception nouns are shown separately. For children, there was no difference between exception and majority-particle nouns. For adults, noun-type and type-frequency interacted, with poorer performance on exception compared to majority-particle nouns in the high type-frequency condition only.

## 2AFC Test

**Children:** The full model failed to converge – we found a simplified converging model by removing the random by-participants slope for the interaction between type-frequency and noun-type, and the correlations between slopes. Relevant means are shown in Figure C2. As in the previous model, the effect of type-frequency was *ns*. There was also no

significant difference between the exception and majority-particle nouns ( $\beta = 0.41$ ,  $SE = 0.34$ ,  $z = 1.20$ ,  $p = .23$ ) and no interaction with type-frequency.

**Adults:** The full model failed to converge – we found a simplified converging model by removing the random by-participants slope for the interaction between type-frequency and noun-type, and the correlations between slopes. Relevant means are shown in Figure C2. As in the previous model, the effect of type-frequency was *ns*. There was also no significant difference between the two noun-types and no interaction with type-frequency.

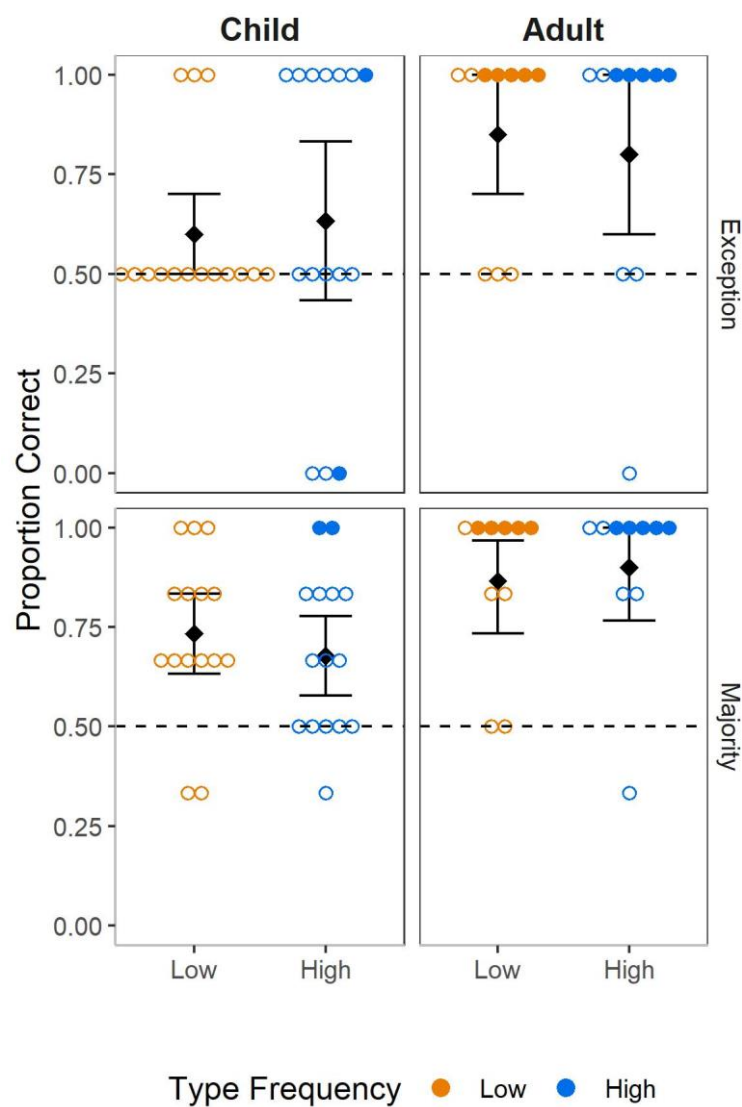


Figure C2: Proportion of correct choices for trained nouns in the 2AFC test for trained nouns, partially consistent condition only, split by noun-type (majority-particle nouns versus exception nouns). Diamonds and

error bars indicate means and bootstrapped 95% confidence intervals. Filled points represent aware participants, unfilled points represent unaware participants. Dashed lines show chance performance. There were no differences between exception and majority-particle nouns in either age group.

### **Production data, split by awareness**

In the analysis of the adult production data set above, performance was higher with nouns occurring with the majority-particle for their semantic class, compared with exception items which occurred with the minority particle, although this appeared to occur primarily in Session 1 and in the high type-frequency condition.

Recall that in the main analyses, we did not see any evidence of semantic learning in participants who could not explicitly describe the relevant patterns (*unaware* participants). However, Figure C1 suggests that the pattern of greater performance with majority-particle than exception nouns in the high type-frequency is seen for both *aware* and *unaware* participants. This is corroborated statistically: analyses on the subset of unaware participants revealed again a main effect of noun-type ( $\beta = 0.72$ ,  $SE = 0.26$ ,  $z = 2.79$ ,  $p = .005$ ), qualified by an interaction with type-frequency ( $\beta = 1.98$ ,  $SE = 0.52$ ,  $z = 3.83$ ,  $p < .001$ ). Noun-type by session and noun-type by type-frequency by session were *ns*. Breaking down the noun-type by type-frequency interaction, there was significantly greater performance on the majority-particle nouns compared to the exception nouns in the high type-frequency condition ( $\beta = 1.751$   $SE = 0.36$ ,  $z = 4.70$ ,  $p < 0.001$ ) but no evidence for this in the low type-frequency condition ( $\beta = -0.27$ ,  $SE = 0.37$ ,  $z = -0.74$ ,  $p = .46$ ), in line with the analyses over all participants. We therefore see some evidence of sensitivity to the semantic cue in adult participants who did *not* verbalize that knowledge in the high type-frequency version of the partially consistent condition.

### **Discussion of Appendix C results**

This analysis revealed that adult participants in the partially consistent condition were more likely to produce incorrect particles with *exception* nouns than *majority particle* nouns

(an effect of noun-type), suggesting a form of over-generalisation on the basis of interference from the semantic cue. This indicates semantic learning and generalisation and is consistent with the findings of the novel nouns tests for adult participants in this condition. However, this difference was only clearly evident in Session 1, and in the high type-frequency condition. The lack of effect of noun-type in Session 4 is likely due to ceiling effects in the learning of trained items, including exception items. However, the interaction with type-frequency is interesting since it could indicate – for the first time in our data – evidence (of a benefit of type-frequency for generalisation (i.e. greater semantic learning when the semantic pattern is exemplified in more lexical items). Unfortunately, there is a caveat: The exception item has higher token frequency the low type-frequency condition (due to each noun being repeated twice as often). An alternative explanation is therefore that semantic learning is actually equivalent in the two type-frequency conditions, but that interference is only seen when the exception item is low in token frequency (i.e. it occurs fewer times during training) and when there are a greater number of items exemplifying the semantic pattern. These explanations could be teased apart in future work by comparing conditions where the type-frequency of the majority category is manipulated, while the exception items are matched in frequency.

Interestingly, this effect of noun-type and interaction with type-frequency was also present when we excluded from our analyses participants who verbally reported the majority semantic patterns at test. This is notable as the only place in our data where we see evidence of semantic generalisation in participants who did not later verbalize that knowledge, suggesting implicit learning. It is interesting that this test, which taps competition between item-level and semantic-class level patterns at an early stage of learning, may better capture effects implicit learning than our novel nouns test. It is also intriguing that this group also

show the type-frequency effect, given that in the General Discussion we noted that type-frequency effects are predicted more under models of implicit rather than explicit learning.

*Table 1.* Participant details and mean standard memory scores (with standard deviations) on the Automated Working Memory Assessment (Alloway, 2008) in each condition.

<b>Age</b>	<b>Consistency</b>	<b>Type-frequency</b>	<b>Mean Age</b>	<b>N males</b>	<b>Verbal Short-Term Memory<sup>1</sup></b>	<b>Visual Short-Term Memory<sup>2</sup></b>	<b>Verbal Working Memory<sup>3</sup></b>
Child	Fully consistent	High	6;0 (0;5)	4/15	113 (13)	112 (11)	108 (9)
		Low	5;11 (0;5)	5/15	111 (14)	118 (16)	106 (28)
	Partially consistent	High	6;0 (0;4)	7/15	102 (17)	113 (18)	106 (20)
		Low	6;2 (0;5)	7/15	102 (13)	117 (13)	113 (13)
	Inconsistent	High	5;11 (0;5)	6/15	102 (13)	107 (16)	108 (17)
		Low	5;11 (0;5)	5/15	105 (16)	109 (13)	110 (20)
Adult	Fully consistent	High	21;3 (5;10)	3/10	100 (12)	95 (23)	99 (20)
		Low	19;3 (0;9)	1/10	104 (22)	101 (17)	102 (19)
	Partially consistent	High	19;1 (0;9)	2/10	100 (19)	101 (14)	98 (6)
		Low	18;10 (0;4)	2/10	101 (18)	91 (11)	103 (6)
	Inconsistent	High	19;4 (0;7)	4/10	94 (17)	89 (12)	109 (17)
		Low	18;10 (0;4)	1/10	104 (18)	94 (14)	104 (6)

<sup>1</sup> Word Recall Task; <sup>2</sup> Maze Memory Task; <sup>3</sup> Backwards Digit Recall Task



*Table 2.* Structure of the language used during sentence training in each experimental condition.

<b>Semantic Consistency</b>	<b>Type-frequency</b>	<b>Particle 1</b>	<b>Particle 2</b>	<b>N repetitions per noun in each training set</b>
Fully consistent	Low	4 animals	4 vehicles	8
	High	8 animals	8 vehicles	4
Partially consistent	Low	3 animals	3 vehicles	8
		1 vehicle	1 animal	
	High	7 animals	7 vehicles	4
		1 vehicle	1 animal	
Inconsistent	Low	2 animals	2 animals	8
		2 vehicles	2 vehicles	
	High	4 animals	4 animals	4
		4 vehicles	4 vehicles	

*Table 3.* Tasks completed in each of the four experimental sessions

<b>Session 1</b>	<b>Session 2</b>	<b>Session 3</b>	<b>Session 4</b>
1) Noun Practice	1) Noun Practice	1) Noun Practice	1) Noun Practice
2) Sentence Training	2) Sentence Training	2) Sentence Training	2) Sentence Training
3) Production Test	3) Verbal Short	3) Visual Short	3) Production Test
	Term Memory Task	Term Memory Task	4) Two Alternative
	4) Sentence Training	4) Sentence Training	Forced Choice Test
		5) Verbal Working	5) Post-Experiment
		Memory Task	Interview

**Table 4:** Summary of core inferential statistics across the paper. Note that all p-values two tailed but where significant indicate effects in the predicted direction, except that significant type-frequency for trained nouns are always in the direction of a higher performance in the low type-frequency condition. Greyed out areas indicate tests which are not part of the design (there is no 2AFC data in session 1) or tests not conducted.

	Trained Nouns				Novel Nouns			
	Children		Adults		Children		Adults	
	Production	2AFC	Production	2AFC	Production	2AFC	Production	2AFC
Overall intercept (comparison to chance)	p < .001	p < .001	p < .001	p < .001				
Intercept, Fully consistent, Session 1					p = .008 <sup>1</sup>		p < .001	
Intercept, Fully consistent, Session 4					p < .001 <sup>1</sup>	p < .001 <sup>1</sup>	p < .001	p < .001
Intercept, Partially consistent, Session 1					ns <sup>2</sup>		p = .004 <sup>1</sup>	
Intercept, Partially consistent, Session 4					ns <sup>2</sup>	ns <sup>2</sup>	p < .001 <sup>1</sup>	p < .002 <sup>1</sup>
Type-frequency	ns	ns	p = .042	ns	ns <sup>3</sup>	ns <sup>3</sup>	ns <sup>3</sup>	ns <sup>3</sup>
Session	p < .001		p < .001		p < .001		p = .03	
Type-frequency * Session	ns				ns		ns	
Fully consistent vs. Partially consistent	p = .002	p = .02	p = .001	ns	p = .002	p = .001	p = .001	p = .005
Fully consistent vs. Partially consistent * Type-frequency	ns	ns	ns	ns	ns	ns	ns	ns
Fully consistent vs. Partially consistent * Session	p < .001		ns		p = .007		ns	
Fully consistent vs. Partially consistent * Type-frequency * Session	ns		ns		ns		ns	
Partially consistent vs. Inconsistent	ns	ns	ns	ns				
Partially consistent vs. Inconsistent * Type-frequency	ns	ns	ns	ns				
Partially consistent vs. Inconsistent * Session	ns		ns					
Partially consistent vs. Inconsistent * Type-frequency * Session	p = .041		ns					

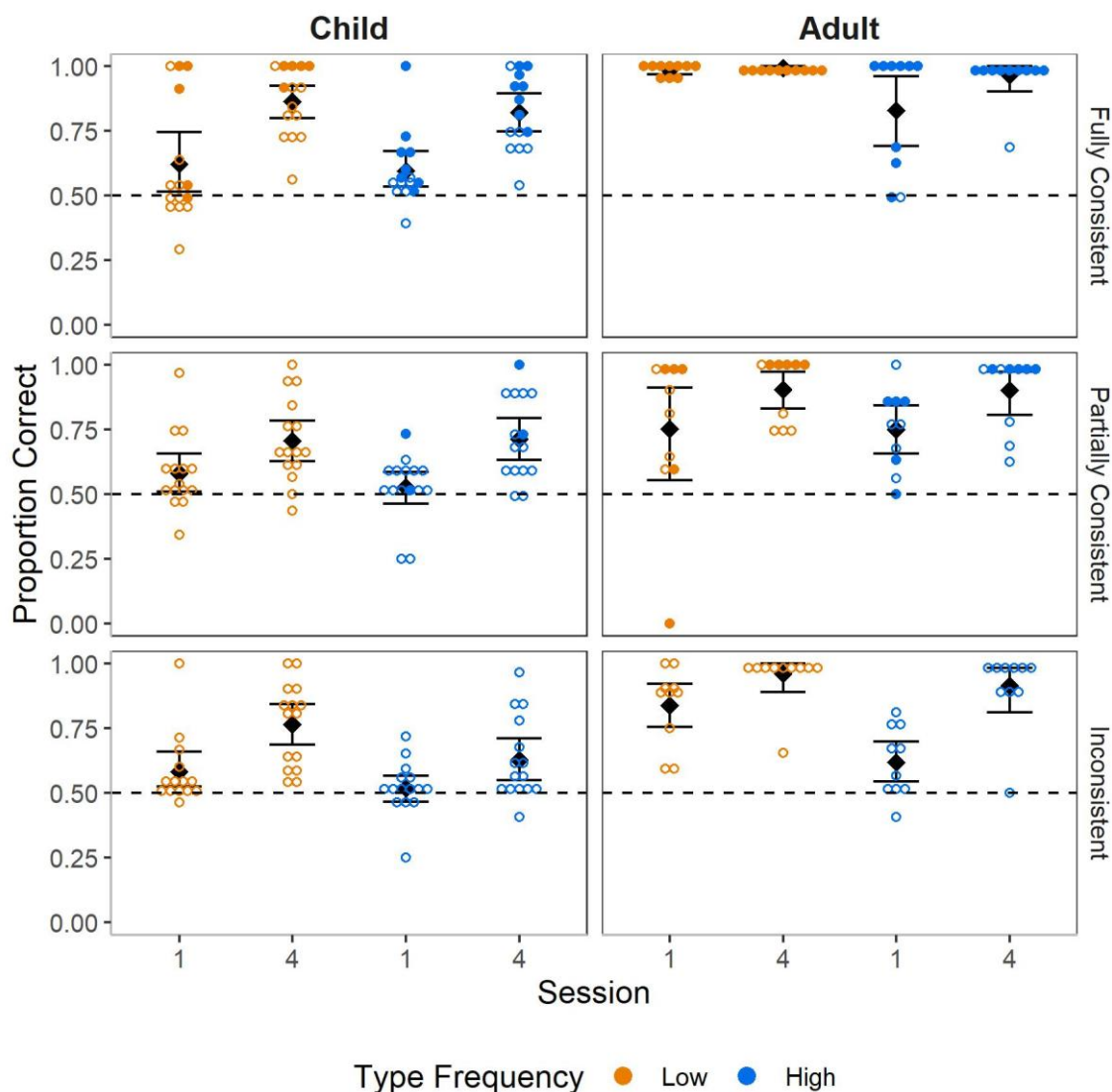
<sup>1</sup> Once “aware” participants were excluded, tests of intercepts against chance for novel nouns where numbers permitted (children – fully consistent; adult – partially consistent) become *ns*, with follow up Bayes Factor analyses suggesting evidence for the null in children and ambiguous evidence in adults

<sup>2</sup> Follow up Bayes Factor analyses suggest that evidence for children in the partially consistent condition is ambiguous

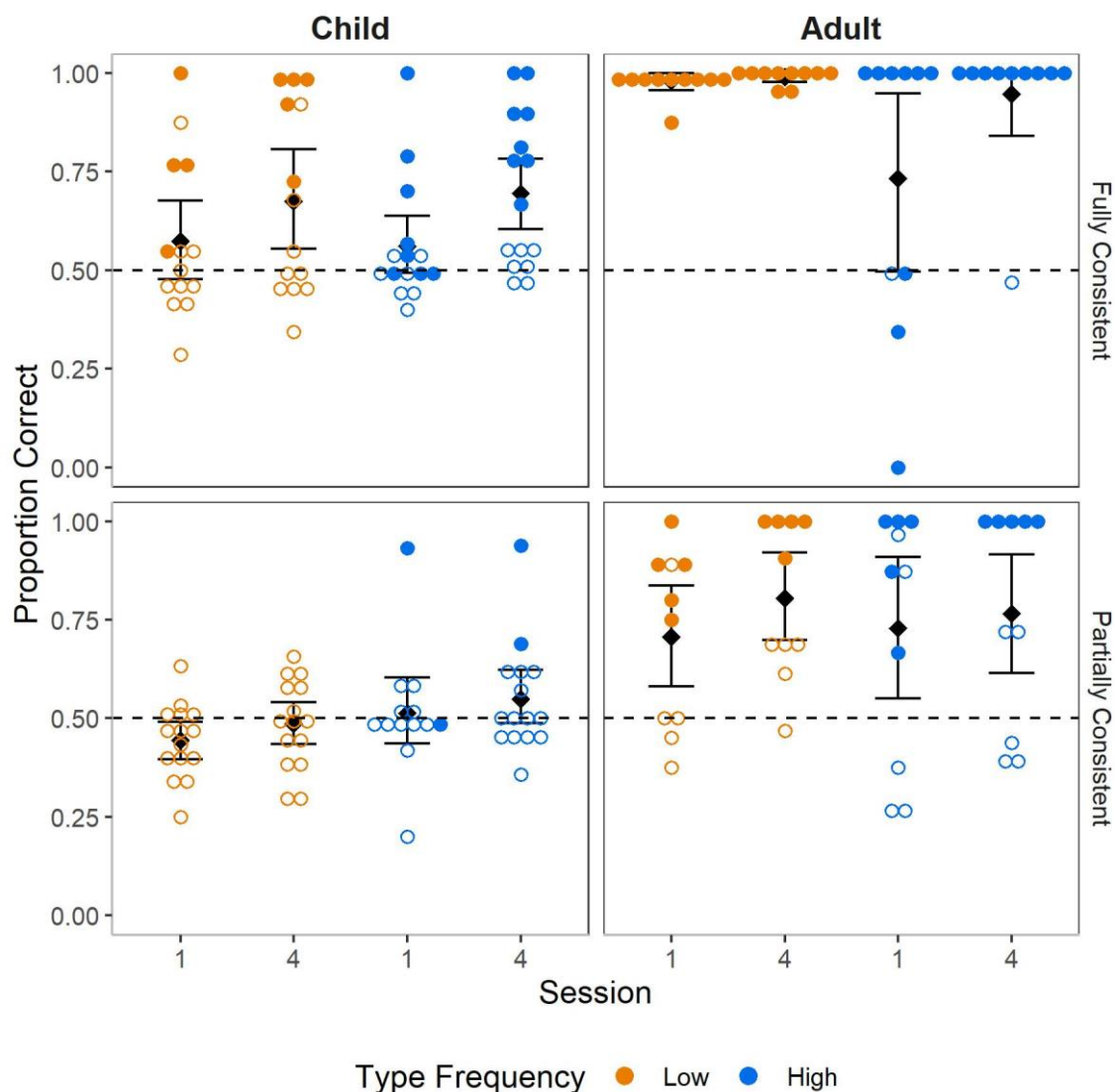
<sup>3</sup> Follow up Bayes Factor analyses suggest evidence for the null for adults and ambiguous evidence for children (analyses for children were conducted on the fully consistent condition only, since that is where children showed generalisation)

*Table 5.* Aware/unaware participants in each experimental condition.

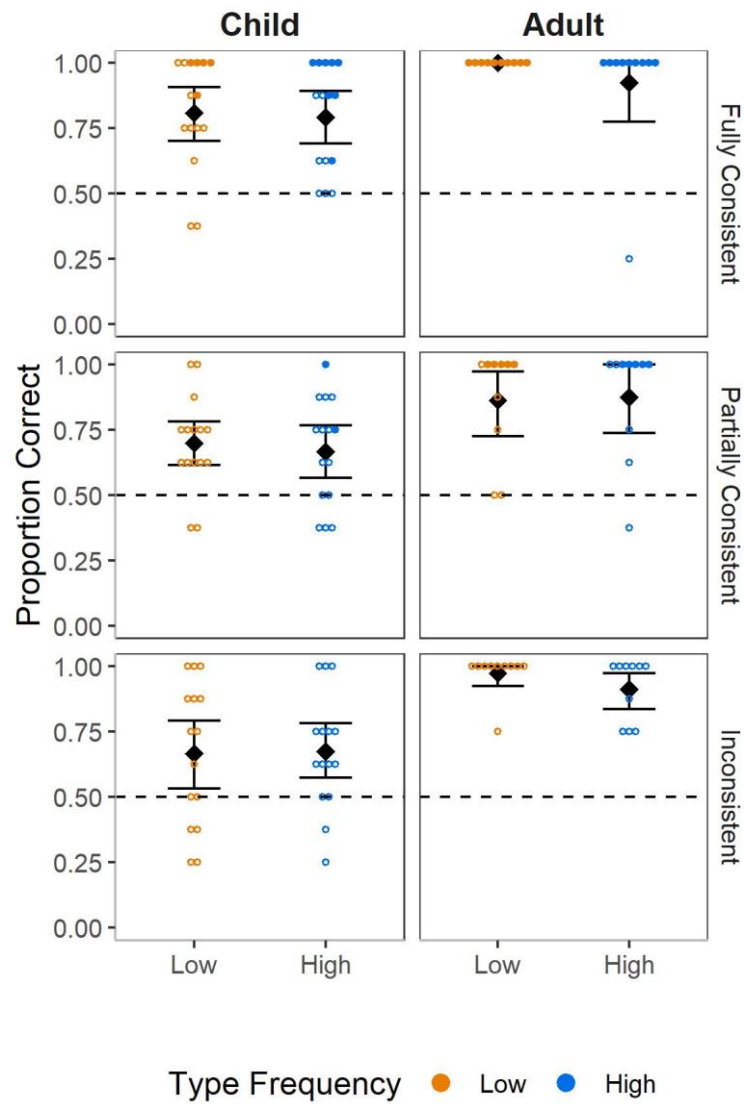
Consistency	Type-frequency	Children		Adults	
		<i>Aware</i>	<i>Unaware</i>	<i>Aware</i>	<i>Unaware</i>
Fully Consistent	High	8	7	9	1
	Low	5	10	10	0
Partially Consistent	High	2	13	5	5
	Low	0	15	5	5



*Figure 1:* Proportion of correctly produced particles for trained nouns. Diamonds and error bars indicate means and bootstrapped 95% confidence intervals. Each coloured point represents a single participant: filled points represent participants who reported the association between particle use and the semantic categories in the post-experiment interview, unfilled points represent participants who were apparently unaware of this association. Dashed lines show chance performance. Performance increased over time in both age groups. Performance was higher in the fully consistent condition than the other two conditions, with no difference between the partially consistent and inconsistent conditions in either age group. Notably, this pattern was present in both sessions for adults, but emerged only in Session 4 for children. Only adults showed a (reverse) type-frequency effect, with higher performance in the smaller, low type-frequency condition.



*Figure 2:* Proportion of correctly produced particles for novel nouns. Diamonds and error bars indicate means and bootstrapped 95% confidence intervals. Each coloured point represents a single participant: as in Figure 1, filled points represent participants who reported awareness of the association between particle use and the semantic categories, unfilled points represent participants who were unaware. Dashed line shows chance performance. For children, accuracy increased over time only in the fully consistent condition. Performance in the partially consistent condition did not exceed chance in either session. For adults, accuracy increased over time in both conditions, and was higher in the fully consistent condition in both sessions, although performance in the partially consistent condition was above chance at both time points. There was no effect of type-frequency for either age group.



*Figure 3:* Proportion of correct choices for trained nouns in the 2AFC test. Diamonds and error bars indicate means and bootstrapped 95% confidence intervals; coloured point represent a single participant, filled points represent aware participants, unfilled points represent unaware participants. Dashed line shows chance performance. For children, performance was higher in the fully consistent condition than the partially consistent and inconsistent conditions. There was no difference between the latter two conditions. For adults, there were no significant differences between any of the semantic consistency conditions. There was no type-frequency effect for either age group.

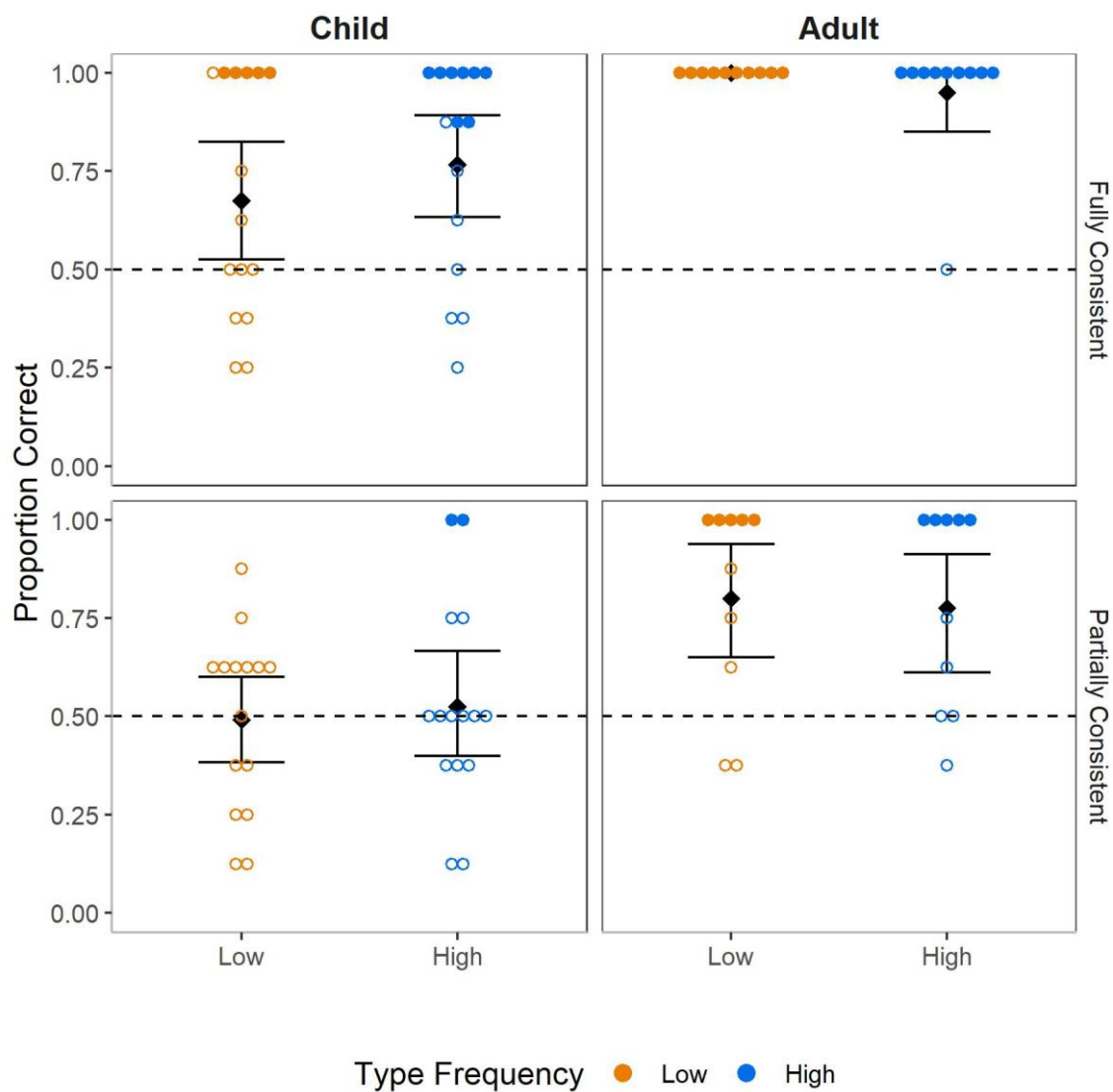


Figure 4: Proportion of correct choices for novel nouns in the 2AFC test. Diamonds and error bars indicate means and bootstrapped 95% confidence intervals; coloured points represent a single participant, filled points represent aware participants, unfilled points represent unaware participants. Dashed lines represent chance performance. Both age groups showed higher performance in the fully consistent condition. Only adults showed above-chance performance in the partially consistent condition. There were no type-frequency effects.