HHS Public Access

Author manuscript

J Child Lang. Author manuscript; available in PMC 2018 May 01.

Published in final edited form as:

J Child Lang. 2018 May ; 45(3): 717–735. doi:10.1017/S0305000917000435.

Distributional learning aids linguistic category formation in school-age children*

Jessica HALL,
University of Iowa, Iowa City, Iowa, USA

Amanda OWEN VAN HORNE, and
University of Delaware, Newark, Delaware, USA

Thomas FARMER
University of Iowa, Iowa City, Iowa, USA

Abstract

The goal of this study was to determine if typically developing children could form grammatical categories from distributional information alone. Twenty-seven children aged six to nine listened to an artificial grammar which contained strategic gaps in its distribution. At test, we compared how children rated novel sentences that fit the grammar to sentences that were ungrammatical. Sentences could be distinguished only through the formation of categories of words with shared distributional properties. Children's ratings revealed that they could discriminate grammatical and ungrammatical sentences. These data lend support to the hypothesis that distributional learning is a potential mechanism for learning grammatical categories in a first language.

Introduction

School-age children may not know exactly what a verb is, but their ability to use verbs in novel contexts demonstrates that they have learned the rules by which verbs operate. An important question in language acquisition is how children learn the grammatical category of a word. In languages with consistent word order, the lexical context in which a word appears can be an important cue, especially in combination with other cues including semantics and phonology (Farmer, Christensen, & Monaghan, 2006; Lany & Saffran, 2011; Monaghan, Christiansen, & Chater, 2007). Here, we aim to determine whether children can learn about grammatical category membership solely from distributional regularities in an artificial language.

Distributional information is distinct from surface-level adjacent and non-adjacent dependencies because it involves tracking these sequential statistics across time and

*We thank Elizabeth Wonnacott for her assistance with Bayesian analyses, and found Wonnacott, Nation, and Brown (2017) particularly helpful for reporting and interpreting Bayes factors. We also thank Tim Arbisi-Kelm, Caitie Hilliard, Sarah O'Neill, and Elissa Newport for their help with stimuli creation. Research reported in this publication was supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under Award Number F31DC015370. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Address for correspondence: Jessica Hall, Communication Sciences & Disorders, University of Iowa, 250 Hawkins IA, Iowa City, Iowa 52240, United States. jessica-e-hall@uiowa.edu.

exemplars. There are limitations to the usefulness of sequence-based types of information for category learning. Despite positive findings from corpus analyses and models using adjacent dependencies to categorize content words (Mintz, Newport, & Bever, 2002; Redington, Chater, & Finch, 1998; Thorpe & Fernald, 2006), adjacent dependencies alone are likely not as helpful for learning functional grammatical categories such as prepositions and conjunctions, which might precede or follow any number of words or word types. Non-adjacent dependencies, captured as frequent frames like "I __ it", have been shown to facilitate word learning in 30-month-olds (Childers & Tomasello, 2001) and reliably contain words of the same grammatical category (Mintz, 2003). Infants as young as 18 months can track such non-adjacent dependencies, with higher variability of the intervening words resulting in better learning (Gómez, 2002). However, frequent frames like "I __ it", while highly accurate in predicting the category of the intervening word, account for a very small percentage of what children hear (St Clair, Monaghan, & Christiansen, 2010). An experiment which shows that distributional information alone can cue child learners to category membership is an essential step in determining its role in language development.

We wanted to study distributional learning in children because they are faced with the task of discovering grammatical categories. It cannot be assumed that children will perform like adults, given that children have different cognitive and meta-cognitive skills and less experience with language. If children cannot do the task, it limits the conclusions we can draw from adult studies. Furthermore, if one takes seriously the idea that language is formed from exposure to input, school-age children have much less cumulative language exposure than adults do and thus much less experience and skill at extracting categories and subcategories from spoken language. Knowledge of argument structure is an area of language arguably still developing into adolescence (Ambridge, Pine, Rowland, & Chang, 2012) and potentially impaired in children with language learning impairments (Ebbels, 2005).

It was once believed that people could not use distributional cues alone to discover grammatical categories. In studies by Smith (1966, 1969), adults could use the fact that an item occurred first or last in a sequence to determine categories in an artificial language, but were unable to utilize distributional cues to restrict their generalizations to exclude ungrammatical co-occurrence violations (an example in real language of an ungrammatical co-occurrence is "he poured the jar with water" – a verb like "fill" is grammatical, but "pour" is not). In later studies, adults were successful only when morphological markers denoting category membership were present (Frigo & McDonald, 1998), as was the case with infants in a Russian gender paradigm (Gerken, Wilson, & Lewis, 2005). Thus, it seemed that only very salient cues like absolute position or a combination of cues would be sufficient for grammatical category learning. Results from modeling and novel word learning studies supported this hypothesis. A model by Monaghan, Chater, and Christensen (2005) showed that phonological and distributional cues used in combination resulted in the most accurate category assignment for 5,000 frequent nouns and verbs in English (between 65–80%, depending on frequency, with about equal accuracy for nouns and verbs), while either type of cue alone produced much poorer discrimination (between 50–85% for distributional cues alone, with better categorization of nouns than verbs; and between 60–65% for phonological cues alone, with better categorization of verbs than nouns). In a series

of behavioral studies, children and adults used novel verbs in new contexts based on both the distributional and semantic properties of other verbs they heard in those contexts (Ambridge & Lieven, 2011; Ambridge, Pine, & Rowland, 2011, 2012).

In an important next step for determining the role of distributional information in language acquisition, Reeder, Newport, and Aslin (2013) showed that adults could use distributional information in the absence of other cues, including position effects, to construct grammatical-like categories in an artificial grammar learning task. In a series of experiments examining how differing degrees of distributional overlap influenced learning, they demonstrated that adults generalized categories without multiple cues to category membership, and that they restricted their generalizations based on the distributional patterns and amount of exposure. In these experiments, grammatical test items consisted of three-word AXB sequences heard during training, novel AXB sequences which had not appeared together in a trigram in training, and ungrammatical sequences containing a component bigram such as XA or BX that violated the linear order of items heard in training. In Experiment 3, the training contained strategic gaps, such that some test items contained grammatically possible bigrams not heard during training. Because of the shared contexts in the distribution of the words in the grammar, participants could form grammatical categories to determine that novel grammatical sentences were possible while the ungrammatical were not. Figure 1 shows how combinations heard during training (left) could lead to the induction of a category, which would lead to the acceptance of a novel combination (right). That participants rated novel grammatical combinations higher than ungrammatical combinations showed that adults can form rudimentary grammatical categories based on distributional information alone. Because adults rated novel items lower in Experiment 3 than in experiments without distributional gaps, the authors concluded that, while the adults formed categories, they also must have found the gaps meaningful. Thus the differences in distributional information in Experiment 3 led to restricted generalizations, which shows the importance of the shape of distributional information in the formation of categories. Further, because during training participants heard AXB combinations as part of longer strings that optionally contained other categories of words at the beginning and end, categories could not have been determined through position effects, as in Smith (1966, 1969). The study did not include ungrammatical items that contained co-occurrence violations only, and so the ability to use co-occurrence information to form and restrict grammatical categories has remained untested. However, that participants could use distributional information alone to determine category membership suggests that this is a strong mechanism for learning language.

As Reeder *et al.* (2013, p. 52) state in their discussion, “Linguistic input to young language learners likely involves many words with partially overlapping contexts (as in Experiment 3)”. Experiment 3 provides an ideal task for an initial foray into assessing distributional learning ability in children because of the realistic nature of the artificial language: in real language, children also must infer and restrict categories from gaps in the input, never being exposed to the entire corpus. Manipulating the input over several experiments is one way to learn how people form and limit generalizations. It is also possible to compare ratings for different types of items within the same experiment. In the present study, we employ an artificial grammar similar to that in Experiment 3 of Reeder *et al.*, modified for use with children. Test items include both linear order and co-occurrence violations to compare

graded degrees of generalization. It has not been shown whether children can use category co-occurrence alone (as distinct from adjacent dependency information, which is individual item co-occurrence) as a cue to category membership, but theories of language learning often assume this skill (e.g., Tomasello, 2000, 2003). Results from this task will provide a basis for future work examining exposure manipulations such as those in Reeder *et al.* (2013), so that we can understand how children with typical and impaired language development create and restrict generalizations.

Methods

Participants—We recruited 27 typically developing participants, aged 6;0–9;11 (M = 8·4, SD=1·1), 16 of whom were females. An additional eight participants (four females) were later added to the sample. Since these participants were not part of our original planned sample, their data were only included for analyses in which we employ Bayesian Analyses and compute Bayes Factors (since, in contrast to the interpretation of *p*-values in frequentist analyses, Bayes Factors remain a valid measure of evidence even with optional stopping; Dienes, 2016; Rouder, 2014).[1] Participants had normal hearing, scored above 85 (M = 115·7, SD=12·6) on the *Kaufman Brief Intelligence Test Matrices subtest, 2nd edition* (Kaufman & Kaufman, 2004), and had no history of neurological disorders or of receiving speech/language therapy per parent report. In addition, all children completed the *Peabody Picture Vocabulary Test, 4th edition (PPVT-4*; Dunn & Dunn, 2007) to document vocabulary skills (raw: M= 158·3, SD=17·4; standard: M= 120·3, SD=12·5). Pilot work suggested that typical children under six could not reliably complete the task.

Stimuli—We take our artificial grammar from the Reeder *et al.* (2013) study. The grammar consisted of five arbitrary category types: Q, A, X, B, and R, controlled for co-occurrence frequency. Each category contained two or three pseudo-words, e.g., category Q words were *klidum* and *spad*. Training sentences were combinations of (Q)AXB(R), in that order, so that each sentence minimally contained AXB with Q and R words added optionally to avoid position effects.

Training—Participants heard 36 sentences, constructed from 12 of the 27 possible AXB combinations of the language. Sentences were chosen such that only a subset of A's appeared with each X, and a subset of X's appeared with each B. The 12 AXB combinations appeared in three of four possible sentence types, QAXB, AXB, AXBR, or QAXBR, and this corpus of 36 sentences was heard three times for a total of 108 trials. The 'Appendix' lists all training items by AXB combination. During training, while children listened to 'aliens' on a computer saying the sentences, they completed a one-back task, indicating by button press if they heard the current sentence immediately prior. Only data from individuals who scored better than 60% on the one-back task were retained; all participants met this criterion. The one-back task ensured children attended during training.

[1]All analyses were additionally run with this larger dataset, and the pattern of significant findings was the same.

Test—The test consisted of 54 AXB sentences of three types: (1) nine 'grammatical' sentences heard during training, (2) nine novel 'grammatical' sentences, i.e., sentences that were not heard during training but were consistent with the grammar of the artificial language, and (3) 18 'ungrammatical' sentences. Grammatical sentences were heard twice at test; ungrammatical were each heard once to avoid familiarization. For the ungrammatical items, we included three of each combination: AXA, BXB, BXA, XBA, AXR, and QXB. AXR and QXB represent co-occurrence violations, while the others represent linear order violations. In each test trial, children listened to a sentence, then chose one of two buttons to indicate whether the sentence was something the alien would say, and finally, they moved a slider scale to indicate confidence (see Ambridge, Pine, Rowland, & Young, 2008), which provided increased power to detect variability in learning. The 'Appendix' lists all sentences heard during the test.

Procedures—The one-back task and training paradigm were introduced in the context of aliens trying to repair their spaceship. Children were told to listen to see if the aliens repeated themselves and to press the red button if the alien said the same thing she just said, and to press the green button if she said something different. Red and green buttons were stickers over the keyboard buttons D and K. Training trials were videos of an alien 'speaking' one of the training sentences. For every trial except the first, participants performed the one-back task. They were also told that videos would play along the way to alert them of the aliens' progress. Videos were four-second clips that included depictions of the aliens' attempts to fix their ship. These occurred at fixed intervals during training.

Immediately after training, participants performed the test. They were told that the aliens again needed their help, and that this time they had to listen to a sentence and decide if it sounded like something the alien would say. Examiners told participants to press the red button if the sentence did not sound like something the alien would say, and green if it did sound like something the alien would say. Then they had to indicate their certainty on a red- and green-colored slider scale by placing the marker on one extreme or the other if they were sure of their decision, and somewhere in the middle if they were less sure. Participants were trained to use the scale at the outset of the experiment through a separate apple/pear shape and color sorting task. E-Prime was used to deliver all task components and perform data collection.

The grammar used in this study was inspired by Experiment 3 of Reeder *et al.* (2013), but we made several departures from that design. For clarity's sake, we list key differences here: stimuli were recorded in child-directed speech, the training contained 12 rather than nine AXB combinations, such that three trained AXB types never appeared during test, training included watching videos with pauses between sentences for a one-back task rather than listening to a continuous audio stream, the training was shortened to three exposures of each sentence because pilot testing showed that children could learn at this exposure with better attention to the task, ungrammatical items at test included additional item types beyond AXA and BXB, the test included a button press and continuous visual scale rather than a 5-point Likert scale, and the apple/pear task ensured that children could use the buttons and the scale.

Analysis

We used the lme4 package (Bates, Maechler, Bolker, & Walker, 2016) and the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2016) in R version 3.1.3 (R Core Team, 2015) to run linear mixed-effects models to explore several comparisons of interest. We used the maximal random effects structure as recommended in Barr, Levy, Scheepers, and Tily (2013), except in instances in which the Akaike information criterion (AIC) and log-likelihood ratios indicated a reduced model improved fit. We compared ratings at test from both binary and slider scale measures for grammatical to ungrammatical items, familiar to ungrammatical, and novel to ungrammatical, to determine learning within groups. We also ran a post-hoc familiar to novel comparison. Novel to ungrammatical is the critical comparison for formation of categories as it is evidence that participants learned the grammar of the artificial language, and that they did so by forming grammatical categories. Ungrammatical served as the reference category except where otherwise noted. We ran a separate model that included age in months and raw scores from the PPVT-4 (Dunn & Dunn, 2007) as covariates to test for factors that contributed to learning. Accuracy on the one-back test during training was also included as a covariate in the model to determine whether attention during exposure influenced the ability to learn categories. Because attention during training would equally affect all item types, we only included this as a main effect, while vocabulary and age could interact with item type.

Results

No child scored below 60% on the one-back task during training (M = 0·86, SD = 0·10), and so all participants' data were included. A linear mixed-effects model with a random subject slope for item type and random intercepts for subject and item was the maximal random effects structure supported by the data. Log-likelihood ratios and AIC confirmed the maximal effects structure as the preferred model. We first ran a model with item grammaticality as the single predictor. Participants rated grammatical sentences (which included familiar and novel items) as more acceptable than ungrammatical in the visual analog scale [beta = −16·04, SE = 3·29, t(36·77) = −4·87, p < ·0001], providing evidence that they could perform the tasks. Results from the binary choice followed those of the visual analog scale, with slightly smaller p-values, for all findings. As such, we report only visual analog scale results from this point forward.

To test learning, we replaced grammaticality with item type (familiar, novel, ungrammatical) in the model. Familiar and novel items were both rated higher than ungrammatical items [familiar: beta = 17·58, SE=4·02, t (34·55) = 4·37, p < ·0001; novel: beta = 14·75, SE=3·83, t(31·15) = 3·85, p < ·001]. Estimated effect size for novel vs. ungrammatical is calculated by finding the correlation between the fitted and observed values of the model (Xu, 2003) ($\Omega_0^2 = .25$). For the purpose of comparing familiar and novel item ratings, we changed the reference category to familiar, and found that ratings for familiar and novel items did not differ [beta = −2·84, SE=4·18, t(28·08) = −0·68, p = ·50]. The mean rating for familiar items was 66·09 (SE = 1·47), for novel items was 63·26 (SE = 1·55), and for ungrammatical items was 48·51 (SE = 1·63). Figure 2 illustrates mean ratings by item type for each participant.

All participants followed a pattern of numerically higher mean ratings for novel grammatical items than ungrammatical.

We converted scale ratings to *z*-scores as in Reeder *et al.* (2013) to control for variable use of the scale. We use *z*-score ratings as the dependent variable from this point forward. Familiar and novel ratings were still rated higher than ungrammatical [familiar: beta = 0·57, SE=0·12, $t(33{\cdot}69) = 4{\cdot}55$, $p < {\cdot}0001$; novel: beta = $t(32{\cdot}52) = 3{\cdot}97$, $p < {\cdot}001$, $\Omega_0^2 = .14$. Familiar and novel ratings did not differ [beta = −0·09, SE=0·13, $t(28{\cdot}30) = -0{\cdot}69$, $p = {\cdot}50$]. Single sample *t*-tests confirmed that mean ratings for all item types were different from zero [familiar: $t(485) = 5{\cdot}31$, $p < {\cdot}0001$; novel: $t(485) = 3{\cdot}03$, $p = {\cdot}003$; ungrammatical: $t(485) = -7{\cdot}43$, $p < {\cdot}0001$].

To determine whether participants showed graded effects of generalization for different ungrammatical item types, we re-ran the mixed effects model first with ungrammatical items with co-occurrence violations excluded and then with items that violated linear order excluded. Our key question was whether the data provided evidence of distributional learning from co-occurrence information alone. Since we added the additional eight participants for this analysis (see ‘Participants’ section), the key inferential statistic computed here is a Bayes Factor for the critical comparison of novel items to co-occurrence violations. We nevertheless also report the *p*-values associated with the coefficients of the mixed model, although their interpretation is limited by the fact that we increased the number of participants from the original sample. Bayes Factors compare evidence supporting the null hypothesis of no difference versus the alternative hypothesis of a significant difference. A Bayes Factor smaller than 1/3 is interpreted as evidence for the null hypothesis, whereas a Bayes Factor greater than 3 is interpreted as evidence for the alternative hypothesis, and Bayes Factors between these values are interpreted as insufficient data for the distinction (see Dienes, 2008, 2014). We used the free online Bayes calculator (Dienes, 2008) for these analyses. Because previous experiments have shown participants can detect linear order violations, we used the mean difference between ratings for novel and linear order violations as the estimate of predicted difference, using the estimate from the mixed-effects model comparing ratings for novel versus ungrammatical items, with co-occurrence violation items excluded. This estimate serves as the SD of a half normal distribution, as per Dienes (2008). The mixed-effects model with co-occurrence violation items excluded showed that participants could distinguish between novel items and linear order violations [beta = 0·51, SE=0·12, $t(31{\cdot}14) = 4{\cdot}24$, $p < {\cdot}001$]. For the sample estimate, we used the coefficient from the mixed-methods model comparing ratings for novel versus ungrammatical items with linear order violations excluded. Results from the mixed effects model with linear order violations excluded and the Bayesian analysis suggested that participants could distinguish between novel items and co-occurrence violations [beta = 0·34, SE=0·17, $t(24{\cdot}29) = 2{\cdot}02$, $p = {\cdot}0549$, BF=3·71].

Taking this one step further, we ran an additional model with familiar items removed to compare novel items to each type of ungrammatical item (AXA, BXB, BXA, XBA, QXB, AXR) to determine whether each type of co-occurrence violation (QXB, AXR) could be distinguished from novel items; novel items served as the reference category. For the

Bayesian analysis, we used the mean difference between ratings for novel items and ungrammatical AXA + BXB items, ·44, as the estimate of predicted difference because Reeder *et al.* (2013) used these items, thus allowing us to predict that participants would be able to distinguish between these types. The sample estimate was the coefficient for each item type in the mixed-effects model. For each item type except AXR, the mixed-effects model and the Bayes Factors supported the alternative hypothesis that ratings for novel items exceeded ratings for that ungrammatical item type [BXA: beta = −0·47, SE=0·19, t(22·73) = −2·55, p = ·02, BF=10·33; XBA: beta = −0·71, SE=0·22, t(23·20) = −3·26, p = ·003, BF=57·55; QXB: beta = −0·46, SE=0·19, t(22·73) = −2·48, p = ·02, BF=9·25; AXR: beta = −0·21, SE = 0·19, t(22·73) = −1·13, p = ·27, BF=1·12]. For AXR, the Bayes Factor indicated insubstantial evidence for the null or alternative hypothesis.

Using the original dataset of 27 from this point forward, we added centered raw PPVT-4 scores, one-back accuracy, and age in months as factors in the model. Log-likelihood ratio tests and AIC comparison indicated that only a random item intercept was needed, likely because subject-specific variance was addressed by standardizing the rating scale. The best fit model included only item type and PPVT-4. The main effect of PPVT-4 was not significant [beta = −0·004, SE=0·002, t(1419) = −1·64, p = ·10], but there was an interaction with item type, such that children with higher vocabulary scores showed a larger distinction between familiar and ungrammatical items than children with lower vocabulary scores [beta = 0·008, SE=0·003, t(1419) = 2·44, p = ·02, Ω_0^2 = .14] (see Figure 3). The slope difference between novel and familiar items was not significant [beta = −0·005, SE=0·003, t(1419) = −1·40, p = ·16]. The regression model is reported in Table 1.

As additional evidence that distributional learning, rather than surface-level adjacent dependencies, drove performance in this task, we calculated associative chunk strength for each item, similar to the bigram analysis Reeder *et al.* (2013) performed for their Experiment 3 results. Associative chunk strength is the average of the frequency during training of the three component dependencies of every test item: each bigram (AX and XB) and the trigram as a whole (AXB) (Knowlton & Squire, 1994). For example, the ungrammatical test item *bleggin zub glim* has an associative chunk strength of ‘0’ because *bleggin zub*, *zub glim*, and *bleggin zub glim* are each never heard during training. In contrast, the ungrammatical sentence *bleggin lapal fluggit* has an associative chunk strength of ‘6’ because the bigram *bleggin lapal* occurs zero times, *lapal fluggit* occurs 18 times, and *bleggin lapal fluggit* occurs zero times. The ‘Appendix’ lists associative chunk strength and mean scale rating for each test item. It is possible for ungrammatical and novel items to have the same chunk strength. If participants were using only this information to perform the task, we would expect similar acceptability ratings for items with identical chunk strength. However, as Figure 4 shows, participants’ ratings differ by item type for items with overlapping chunk strength. Adding chunk strength to the model did not reveal any significant effect, and the effect of item type remained significant. Thus it appears that distributional information that goes beyond information derived solely from adjacent dependencies aided participants’ performance in this task. However, ungrammatical items had the widest range of scores (see ‘Appendix’), which suggests that item-level properties beyond chunk strength may have influenced ratings.

Discussion

We used an artificial grammar similar to that in Experiment 3 of Reeder *et al.* (2013) to test distributional learning in typically developing school-age children. This task employed systematic gaps in the exposure such that learners had to formulate grammatical categories based on shared distributions to distinguish novel grammatical from ungrammatical items. In this way, the task simulates the problem of real language learning in that children must distinguish possible combinations from impossible without explicit knowledge of category membership. Children could distinguish grammatical from ungrammatical items in a grammar they were recently exposed to. Critically, they distinguished novel grammatical items from ungrammatical items through the utilization of distributional information in the grammar. They also showed graded performance, with lower ratings for linear violations than co-occurrence violations, though co-occurrence violations were still rated lower than novel grammatical items (at least for QXB items), evidence that children can use this information to inform category membership. Item chunk strength analysis, which is based on how often two words appear next to each other, confirmed that children were not using adjacent dependency frequency alone to perform the task, though the observation of lower ratings for items with lower chunk strength show that children constrain generalizations to those that are more robust statistically. Results show that children as young as six are sensitive to distributional information and can utilize it, even in the absence of other strong cuing information such as semantics or phonology, to form categories.

While different scales make a direct comparison with Reeder *et al.* (2013) untenable, our finding of similar ratings for familiar and novel items is somewhat inconsistent with adult performance in the earlier study. This may be due to a shortened overall exposure (three repetitions of the corpus instead of four), as Reeder *et al.* found lower ratings for novel items in Experiment 4, which had extended exposure time. The lack of a difference in our study suggests participants are not merely using stored sentence strings from the training to perform the test, as familiar ratings would be higher if participants relied on this strategy. These results fit with a theory of category learning, whereby individual exemplars (the sequence *lapal fluggit* or *daffin lapal bleggin*) may be stored temporarily until a threshold is reached for determining a class of items that can appear in certain contexts, at which point individual exemplars are no longer needed and generalization can occur. Differences between ratings for items with co-occurrence violations and novel grammatical items suggest that information about category membership goes beyond linear order. Recall that co-occurrence violations are not just that these two words never appeared next to each other (an adjacent dependency), but that the two CATEGORIES did not. This is a novel finding for any age of participant. Regarding individual differences in statistical learning, raw scores on the PPVT-4 (Dunn & Dunn, 2007) predicted distinction in ratings for familiar and ungrammatical items. There were no main effects or interactions with age or with accuracy on the one-back task during training, suggesting that distributional learning may be developmentally invariant, and only minimal attention is required, or that other factors may allow for advantages. Other work has suggested individual differences in statistical learning are related to language abilities (Kidd, 2012; Misyak & Christensen, 2012; Misyak, Christensen, & Tomblin, 2010; Seigelman & Frost, 2015). Lany and Saffran (2011) found a

relationship between vocabulary ability and use of distributional cues in linguistic input in a word learning task. In their study, infants with large vocabularies learning an artificial grammar generalized semantic categories using distributional cues more than phonological cues, while infants with smaller vocabularies showed the opposite pattern. If vocabulary scores are understood as a proxy measure of general language aptitude, results from the present study, combined with those of Lany and Saffran, provide some evidence that statistical learning is related to language ability. Given the age of the children in our study, the direction of this relationship is not clear. The parameter estimates and effect size for the interaction are small, and it will be interesting to see if a wider range of vocabulary abilities serves to increase this effect. Future studies, including a replication with children with specific language impairment, will attempt to explore this relationship further.

Adult-like metacognitive skills do not appear to be necessary for distributional learning. Children aged six to nine could both generalize and limit generalizations based on what they heard in the input. Findings from this study support the hypothesis that implicit statistical learning is involved in language acquisition in two ways. One, there is a link between language ability and distributional learning ability. Two, because there is no explicit instruction in the task, children do not need to be fully aware that grammatical rules exist before they begin using information to make generalizations. We also saw that they showed gradation in their formation of categories, with higher ratings for items with co-occurrence violations than linear order violations. A future study that directly manipulates the number of items in the lexicon, as well as length of exposure to the artificial grammar, would reveal how learners use and weight different cues for generalization. This would allow exploration of item-level properties not possible to explore with the grammar of the current study.

We provide evidence that children as young as six can use distributional information in novel linguistic input to form grammatical categories, without other cuing information. Evidence that children use categories comes from higher ratings for novel grammatical test items than ungrammatical items containing similar bigram frequencies. That such a powerful learning mechanism is available to young learners strengthens its plausibility as a useful mechanism in language acquisition. This work provides an important foundation for extending the findings through additional studies on subcategory learning, comparison with adults, and comparison with individuals with language impairment. Future work will also explore manipulations of exposure, as in Reeder *et al.* (2013), to determine in more detail how children limit generalizations.

References

Ambridge, B., Lieven, EV. Child language acquisition: contrasting theoretical approaches. Cambridge University Press; 2011.

Ambridge B, Pine JM, Rowland CF. Children use verb semantics to retreat from overgeneralization errors: a novel verb grammaticality judgment study. Cognitive Linguistics. 2011; 22(2):303–23.

Ambridge B, Pine JM, Rowland CF. Semantics versus statistics in the retreat from locative overgeneralization errors. Cognition. 2012; 123(2):260–79. [PubMed: 22325040]

Ambridge B, Pine JM, Rowland CF, Chang F. The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. Language. 2012; 88(1):45–81.

Ambridge B, Pine JM, Rowland CF, Young CR. The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. Cognition. 2008; 106(1):87–129. [PubMed: 17316595]

Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: keep itmaximal. Journal ofMemory andLanguage. 2013; 68:255–78.

Bates, D., Mächler, M., Bolker, B., Walker, S. lme4: linear mixed-effects models using Eigen and S4. R package version 1.1-12. 2016. Online: <http://CRAN.R-project.org/package=lme4>

Childers JB, Tomasello M. The role of pronouns in young children's acquisition of the English transitive construction. Developmental Psychology. 2001; 37(6):739–48. [PubMed: 11699749]

Dienes, Z. Understanding psychology as a science: an introduction to scientific and statistical inference. New York: Palgrave Macmillan; 2008.

Dienes Z. Using Bayes to get the most out of non-significant results. Frontiers in Psychology. 2014; 5:781. [PubMed: 25120503]

Dienes Z. How Bayes factors change scientific practice. Journal of Mathematical Psychology. 2016; 72:78–89.

Dunn, LM., Dunn, DM. *PPVT-4: Peabody picture vocabulary test*. Minneapolis, MN: Pearson Assessments; 2007.

Ebbels, SH. Unpublished doctoral dissertation. University of London; 2005. Argument structure in specific language impairment: from theory to therapy.

Farmer TA, Christiansen MH, Monaghan P. Phonological typicality influences on-line sentence comprehension. Proceedings of the National Academy of Sciences. 2006; 103:12203–08.

Frigo L, McDonald JL. Properties of phonological markers that affect the acquisition of gender-like subclasses. Journal of Memory and Language. 1998; 39(2):218–45.

Gerken L, Wilson R, Lewis W. Infants can use distributional cues to form syntactic categories. Journal of Child Language. 2005; 32(2):249–68. [PubMed: 16045250]

Gómez RL. Variability and detection of invariant structure. Psychological Science. 2002; 13(5):431–6. [PubMed: 12219809]

Kaufman, AS., Kaufman, NL. K-BIT-2: Kaufman brief intelligence test. 2. Circle Pines, MN: American Guidance Service, Inc; 2004.

Kidd E. Implicit statistical learning is directly associated with the acquisition of syntax. Developmental Psychology. 2012; 48(1):171–84. [PubMed: 21967562]

Knowlton BJ, Squire LR. The information acquired during artificial grammar learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1994; 20(1):79–91.

Kuznetsova, A., Brockhoff, PB., Christensen, RHB. lmerTest: tests in linear mixed effects models. R package version 2.0-32. 2016. Online: <https://cran.r-project.org/web/packages/lmerTest/index.html>

Lany J, Saffran JR. Interactions between statistical and semantic information in infant language development. Developmental Science. 2011; 14(5):1207–19. [PubMed: 21884336]

Mintz TH. Frequent frames as a cue for grammatical categories in child directed speech. Cognition. 2003; 90(1):91–117. [PubMed: 14597271]

Mintz TH, Newport EL, Bever TG. The distributional structure of grammatical categories in speech to young children. Cognitive Science. 2002; 26(4):393–424.

Misyak JB, Christiansen MH. Statistical learning and language: an individual differences study. Language Learning. 2012; 62(1):302–31.

Misyak JB, Christiansen MH, Tomblin JB. On-line individual differences in statistical learning predict language processing. Frontiers in Language Sciences. 2010; 1:31.

Monaghan P, Chater N, Christiansen MH. The differential role of phonological and distributional cues in grammatical categorisation. Cognition. 2005; 96(2):143–82. [PubMed: 15925574]

Monaghan P, Christiansen MH, Chater N. The phonological-distributional coherence hypothesis: cross-linguistic evidence in language acquisition. Cognitive Psychology. 2007; 55(4):259–305. [PubMed: 17291481]

Redington M, Chater N, Finch S. Distributional information: a powerful cue for acquiring syntactic categories. Cognitive Science. 1998; 22(4):425–69.

Reeder PA, Newport EL, Aslin RN. From shared contexts to syntactic categories: the role of distributional information in learning linguistic form-classes. Cognitive Psychology. 2013; 66(1): 30–54. [PubMed: 23089290]

Rouder JN. Optional stopping: no problem for Bayesians. Psychonomic Bulletin & Review. 2014; 21:301–8. [PubMed: 24659049]

Siegelman N, Frost R. Statistical learning as an individual ability: theoretical perspectives and empirical evidence. Journal of Memory and Language. 2015; 81:105–20. [PubMed: 25821343]

Smith KH. Grammatical intrusions in the free recall of structured letter pairs. Journal of Verbal Learning and Verbal Behavior. 1966; 5(5):447–54.

Smith KH. Learning co-occurrence restrictions: Rule induction or rote learning? Journal of Verbal Learning and Verbal Behavior. 1969; 8(2):319–21.

St Clair MCS, Monaghan P, Christiansen MH. Learning grammatical categories from distributional cues: flexible frames for language acquisition. Cognition. 2010; 116(3):341–60. [PubMed: 20674613]

Thorpe K, Fernald A. Knowing what a novel word is not: two-year-olds 'listen through' ambiguous adjectives in fluent speech. Cognition. 2006; 100(3):389–433. [PubMed: 16125688]

Tomasello M. The item-based nature of children's early syntactic development. Trends in Cognitive Sciences. 2000; 4:156–63. [PubMed: 10740280]

Tomasello, M. Constructing a language: a usage-based theory of language acquisition. Cambridge, MA: Harvard University Press; 2003.

Wonnacott E, Brown H, Nation K. 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. 2017; 95:36–48.

Xu R. Measuring explained variation in linear mixed effects models. Statistics in Medicine. 2003; 22(22):3527–41. [PubMed: 14601017]

Appendix

List of all training items, listed by AXB combination and sentence type.

AXB1	Sentence type
daffin tomber mawg frag	AXB
klidum daffin tomber mawg gentif	QAXBR
AXB2	
daffin tomber bleggin	AXB
daffin tomber bleggin gentif	AXBR
klidum daffin tomber bleggin	QAXB
spad daffin tomber bleggin gentif	QAXBR
AXB3	
daffin zub fluggit	AXB
daffin zub fluggit gentif	AXBR
spad daffin zub fluggit	QAXB
AXB4	
daffin zub bleggin gentif	AXBR
klidum daffin zub bleggin	QAXB
spad daffin zub bleggin frag	QAXBR
AXB5	

AXB1	**Sentence type**
flairb tomber mawg	AXB
klidum flairb tomber mawg gentif	QAXBR
spad flairb tomber mawg	QAXB
AXB6	
glim zub fluggit	AXB
glim zub fluggit gentif	AXBR
klidum glim zub fluggit	QAXB
AXB7	
glim zub bleggin	AXB
glim zub bleggin frag	AXBR
klidum glim zub bleggin frag	QAXBR
AXB8	
glim lapal fluggit	AXB
klidum glim lapal fluggit	QAXB
spad glim lapal fluggit gentif	QAXBR
AXB9	
glim lapal mawg	AXB
glim lapal mawg gentif	AXBR
spad glim lapal mawg	QAXB
AXB10	
flairb tomber bleggin	AXB
klidum flairb tomber bleggin frag	QAXBR
spad flairb tomber bleggin	QAXB
AXB11	
flairb lapal fluggit	AXB
flairb lapal fluggit frag	AXBR
spad flairb lapal fluggit frag	QAXBR
AXB12	
flairb lapal mawg	AXB
flairb lapal mawg frag	AXBR
klidum flairb lapal mawg	QAXB

List of all test items with chunk strength, organized by item type and mean rating.

Test item	**Chunk strength**	**Sentence type**	**Mean *z*-score rating**	**SD**	**Mean rating**
Familiar	**14·67**		**0·22**	**0·91**	**66·09**
flairb lapal fluggit	15	AXB11	0·59	0·76	77·87
glim lapal mawg	15	AXB9	0·51	0·88	75·04
flairb tomber bleggin	15	AXB10	0·43	0·78	73·22
flairb lapal mawg	15	AXB12	0·34	0·80	70·72
daffin tomber mawg	12	AXB1	0·31	0·80	67·44

Test item	Chunk strength	Sentence type	Mean *z*-score rating	SD	Mean rating
daffin zub fluggit	15	AXB3	0·20	0·92	64·65
glim lapal fluggit	15	AXB8	0·01	0·84	60·94
glim zub bleggin	15	AXB7	−0·19	1·09	54·91
daffin zub bleggin	15	AXB4	−0·21	0·94	50·04
daffin tomber bleggin	appeared during training but not at test				
flairb tomber mawg	appeared during training but not at test				
glim zub fluggit	appeared during training but not at test				
Novel	**4**	**AXB**	**0·13**	**0·93**	**63·26**
flairb tomber fluggit	6		0·62	1·03	78·56
daffin tomber fluggit	6		0·43	0·75	72·04
daffin lapal bleggin	0		0·12	0·93	63·74
glim tomber fluggit	0		0·18	0·91	63·67
glim tomber bleggin	6		0·09	0·74	61·52
daffin lapal mawg	6		0·09	0·86	61·43
flairb zub mawg	0		−0·06	0·93	56·41
glim zub mawg	6		−0·20	1·00	56·17
flairb zub bleggin	6		−0·12	0·96	55·78
Ungrammatical	**4·33**		**−0·35**	**1·03**	**48·51**
spad lapal fluggit	6	QXB†	0·25	0·82	68·59
daffin tomber frag	6	AXR†	0·03	0·99	61·93
flairb zub gentif	6	AXR†	0·04	1·06	58·26
mawg lapal daffin	0	BXA	−0·11	1·12	57·22
bleggin lapal fluggit	6	BXB	−0·06	0·98	56·96
glim lapal frag	6	AXR†	−0·15	1·05	55·89
glim zub daffin	6	AXA	−0·22	0·92	52·85
daffin zub flairb	6	AXA	−0·34	0·89	49·30
bleggin zub glim	0	BXA	−0·37	1·03	49·19
flairb lapal glim	6	AXA	−0·35	0·85	47·37
spad tomber bleggin	6	QXB†	−0·41	0·90	44·96
mawg zub bleggin	0	BXB	−0·45	0·84	43·41
tomber mawg glim	3	XBA	−0·54	1·15	42·89
fluggit tomber flairb	6	BXA	−0·68	1·25	39·33
fluggit tomber mawg	6	BXB	−0·65	0·95	38·30
mawg bleggin flairb	0	BBA*	−0·69	1·07	36·93
zub fluggit daffin	3	XBA	−0·67	1·04	36·04
klidum zub mawg	6	QXB†	−0·90	1·05	33·74

Notes: Means are reported in bold. Familiar items also appeared with Q and R words at the beginning and end during training, and sentence type refers to the AXB combination within, as listed in the earlier part of the 'Appendix';

*BBA occurred as a mistake; this item was intended to be an BXA item;

†indicates a co-occurrence violation, as opposed to a linear order violation.

Fig. 1.

This schematization shows that combinations heard during training can lead to the induction of a category of items that have shared distributions, which allows the listener to generalize a new combination at test.

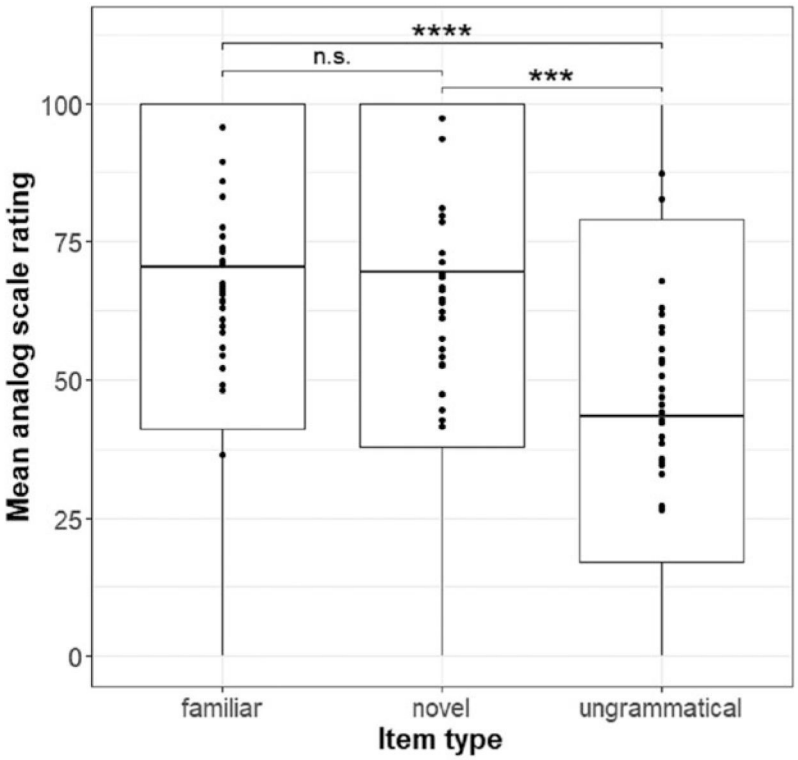

Fig. 2.

Mean scale ratings of test items by item type.

NOTES: *** $p < \cdot 001$, **** $p < \cdot 0001$.

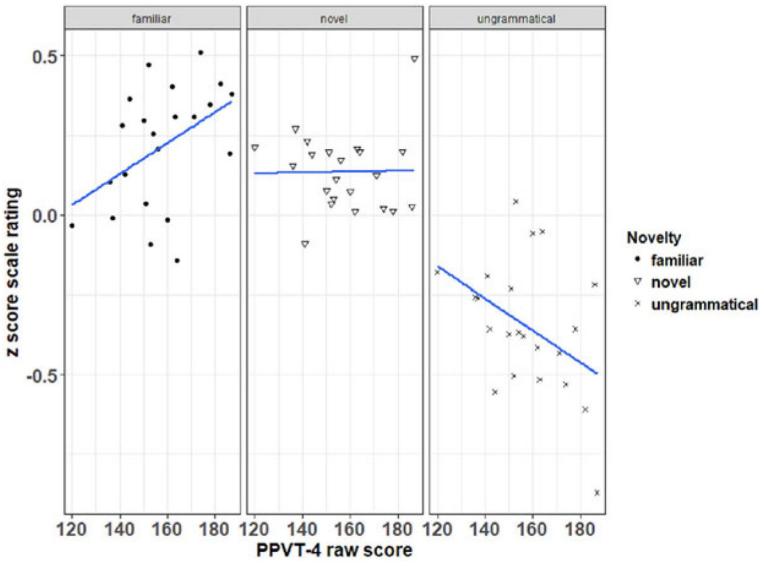

Fig. 3.

Mean *z*-score scale ratings by item type (familiar, novel, ungrammatical) as indexed by the *Peabody Picture Vocabulary Test*, 4th edition (PPVT-4), raw score.

NOTE. PPVT-4 raw scores are not centered for the purpose of illustration, but are centered in the model.

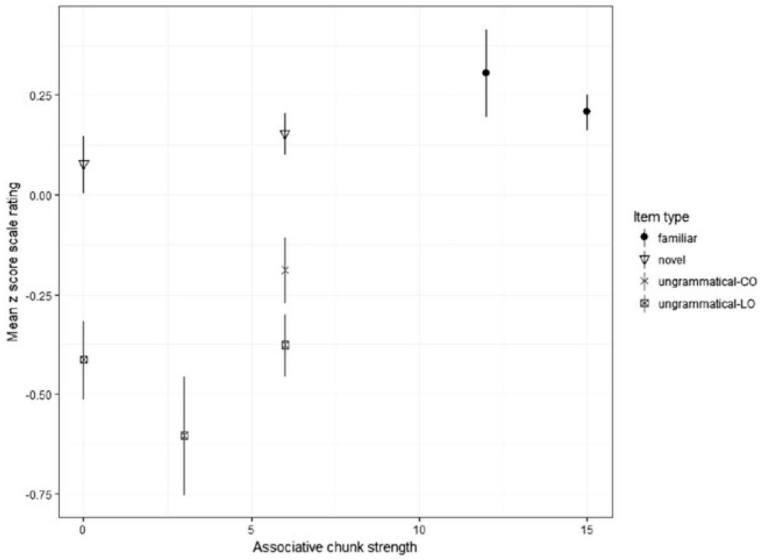

Fig. 4.

Mean *z*-score scale ratings by item type (familiar, novel, ungrammatical co-occurrence violation, ungrammatical linear order violation) and chunk strength.

Table 1

Results of the regression model showing the influence of item type and centered Peabody Pictured Vocabulary Test, 4th Edition (PPVT-4) raw score on ratings

Factor	Variance	*SD*	Coefficient	*SE*	*p*
Random factors					
Item intercept	·06	·25			
Fixed factors (Intercept)			−0·35	0·07	< ·0001
Item type (reference category = ungrammatical)					
Familiar			0·56	0·12	< ·0001
Novel			0·47	0·12	< ·001
PPVT-4 raw score, centered	−0·004	0·002	·10		
PPVT-4 × item type (reference category = ungrammatical)					
Familiar × PPVT-4			0·008	0·003	·02
Novel × PPVT-4			0·004	0·003	·30