Zero-Shot Cross-Situational Learning for Building Word-Referent Mappings

Melina L. Knabe (melina.knabe@utexas.edu)
Department of Psychology, 108 E Dean Keeton Street
Austin, TX 78712 USA

Chen Yu (chen.yu@austin.utexas.edu)
Department of Psychology, 108 E Dean Keeton Street
Austin, TX 78712 USA

Abstract

Statistical learning (SL) mechanisms drive learning across several domains, including language acquisition. Can SL mechanisms like cross-situational word learning, which rely on the accumulation of statistical evidence, also account for the rapid acceleration in children's word learning? This study examined whether learners could track and integrate several statistical regularities concurrently and use this information to map words to objects that never co-occurred during training—a behavior known as zero-shot learning. Experiment 1 showed that learners leveraged their acquired knowledge of word and object categories to map words to objects that did not co-occur during training. Experiment 2 extended this finding by demonstrating that learners integrated their newly acquired statistics with prior knowledge to map novel words to perceptually similar, yet previously unseen, objects. These findings suggest that integrating several types of statistical relationships between words and objects can accelerate new learning, making SL an efficient mechanism for early word learning.

Keywords: statistical learning; cross-situational word learning; zero-shot learning; language acquisition; multimodal statistics

Introduction

Statistical learning (SL) is a powerful mechanism by which species detect and use regularities in their environment (Christiansen, 2019; Santolin & Saffran, 2018; Schapiro & Turk-Browne, 2015). The human learner is fundamentally a statistical learner as SL drives learning across visual, auditory, and language domains (Erickson & Thiessen, 2015; Fiser & Aslin, 2001; Romberg & Saffran, 2010; Saffran et al., 1996). Within the language domain, SL mechanisms have been proposed as solutions to tasks such as word segmentation (Saffran et al., 1996), phonological learning (Maye et al., 2002), syntactic learning (Thompson & Newport, 2007), and word learning (Smith & Yu, 2008; Yu & Smith, 2007).

Cross-situational learning has been studied as a SL solution for word learning (Roembke et al., 2023; Smith & Yu, 2008; Yu & Smith, 2007; Zhang et al., 2019). The standard experimental paradigm of cross-situational learning was designed to test how learners track the co-occurrences of words and objects across several individually ambiguous learning situations to build word-object mappings (Smith & Yu, 2008; Yu & Smith, 2007). This body of work has shown that even young learners can accumulate statistical evidence to learn words in ambiguous learning contexts.

Although accumulating statistical evidence over time can be a slow process, we know that young children are remarkably fast and efficient word learners. By 18-24 months, children show rapid acceleration in their vocabulary growth, acquiring both individual word-object mappings and a broader lexico-semantic network (Goldfield & Reznik, 1990; Plunkett et al., 2022). This raises a central question: Can cross-situational learning, which relies on the accumulation of statistical evidence, also account for the speed and efficiency of real-world word learning?

To address this question, we propose that statistical word learning should be tested in broader and more naturalistic contexts. Whereas cross-situational word learning paradigms narrowly test whether learners can extract word-object co-occurrences, real-world learning environments contain multiple types of statistical regularities. That is, learners not only encounter learning situations with direct word-object co-occurrence statistics, such as hearing "apple" and "banana" while seeing two fruits. They may also hear related words co-occur in speech ("apple", "orange", "pear"), enabling them to build lexical knowledge even without visual referents (Savic et al., 2023; Unger et al., 2020; Willits et al., 2013; Wojcik & Saffran, 2015). Similarly, they may see related objects together in the environment (e.g., a pear on a kitchen counter), allowing them to build semantic knowledge about objects without words (Peters & Borovsky, 2019; Roy et al., 2015; Sadeghi et al., 2015; Tamis-LeMonds et al., 2019; Wojcik & Saffran, 2013).

Our central hypothesis in the present study is that learners can track and integrate several types of statistical relationships among words, objects, and words and objects to accelerate learning. If learners can integrate and use this newly acquired knowledge in subsequent learning, cross-situational learning may become increasingly efficient as more data accumulates (Andrews et al., 2009; Chen et al., 2017; Yu, 2008; Knabe & Vlach, 2023). Crucially, the integration of several types of statistical relationships may even enable learners to build mappings between words and objects that rarely (or never) directly co-occur. The ability to generalize learning to unseen data is known as *few-shot* or *zero-shot learning* and has been well studied in machine learning (Wang et al., 2019; Xian et al., 2017). However, to

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

date, zero-shot learning has not been tested in the context of human statistical word learning.

The Current Study

We tested this hypothesis in two learning contexts (Experiment 1 and Experiment 2), in which learners were first exposed to various statistical relationships among words and objects, and were then asked to identify the referents of words that never co-occurred during training. Because the to-be-learned words and objects never co-occurred together, we call this zero-shot learning of word-object mappings.

Experiment 1 comprised three phases (see Figure 1): In Phase 1, participants learned categories of visual objects and categories of words. Although these categories could be learned from co-occurrences alone, real-world input often contains additional structure. For instance, co-occurring items may also share category membership or visual similarity. Experiment 1 was thus designed so that the categories were based on co-occurrence statistics and shared category membership (e.g., all objects in one category were vehicles). In Phase 2, participants learned the mappings between specific words and objects. This also enabled them to learn how the word and object categories mapped on to each other more broadly (e.g., all vehicles had names with the letter "C"). In Phase 3, participants had to use the individual word-object mappings and overall category mappings to link words and objects that never co-occurred before. Successful zero-shot learning in Phase 3 thus required integrating and leveraging newly acquired statistical relationships from the first two phases.

Experiment 2 examined whether learners could integrate acquired co-occurrences between words and objects with their prior semantic knowledge of visual objects to accelerate new learning (see Figure 2). Participants were exposed to novel words paired with familiar words and/or objects. We then tested whether they could map the novel words to perceptually-similar—but previously unseen—objects.

Together, these experiments demonstrate that learners can integrate multiple types of statistical relationships to support fast and efficient word learning, including zero-shot learning. These findings suggest that statistical learning mechanisms can account for the rapid acceleration in early word learning by making the most of the kind of statistical regularities encountered in the real world.

Method

Experiment 1

The goal of this study was to examine whether learners could use their newly acquired knowledge of visual and word categories to build new word-referent mappings.

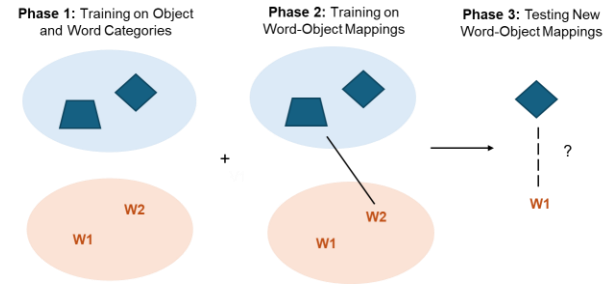

Figure 1: A schematic of Experiment 1. Learners were exposed to categories of visual objects (symbolized by the blue shapes) and words (e.g., W1, W2) in Phase 1. They then learned the direct mappings between several objects and words denoted by the connecting line in Phase 2. Finally, in Phase 3, learners had to integrate their newly acquired knowledge to map words to objects that never co-occurred during training.

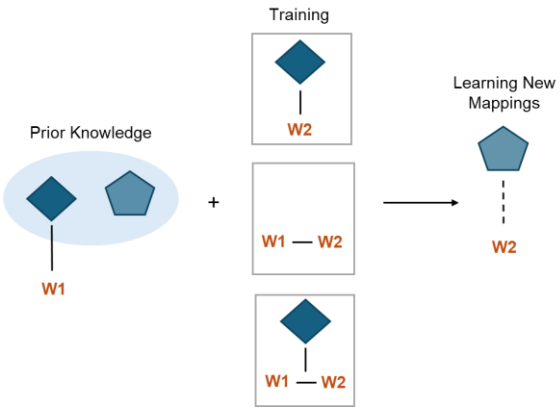

Figure 2: A schematic of Experiment 2. Learners had prior knowledge of the word (W1) for a familiar object and the perceptual similarity between two objects (symbolized by the blue shapes). During training, a novel word (W2) was either paired with a familiar object (Visual-Only), the word for a familiar object (Label-Only), or both (Visual-Label). Learners were then tested on whether they could make an inference to map the novel word to a perceptually-similar object that did not appear during training.

Participants

The participants were 41 adults recruited on the online research platform Prolific (www.prolific.com) and were compensated monetarily for their participation. Participants were on average 37.6 years old (*SD* = 11.7, range: 19-60 years) and included 28 females. A majority (51.2%) identified as White (39%: Black or African American, 7.3%: Asian, 2.4%: declined to respond). A majority (68.2%) reported holding a bachelor's or graduate degree (17.1%: some college, 9.8%: high school degree, 5%: associate's degree).

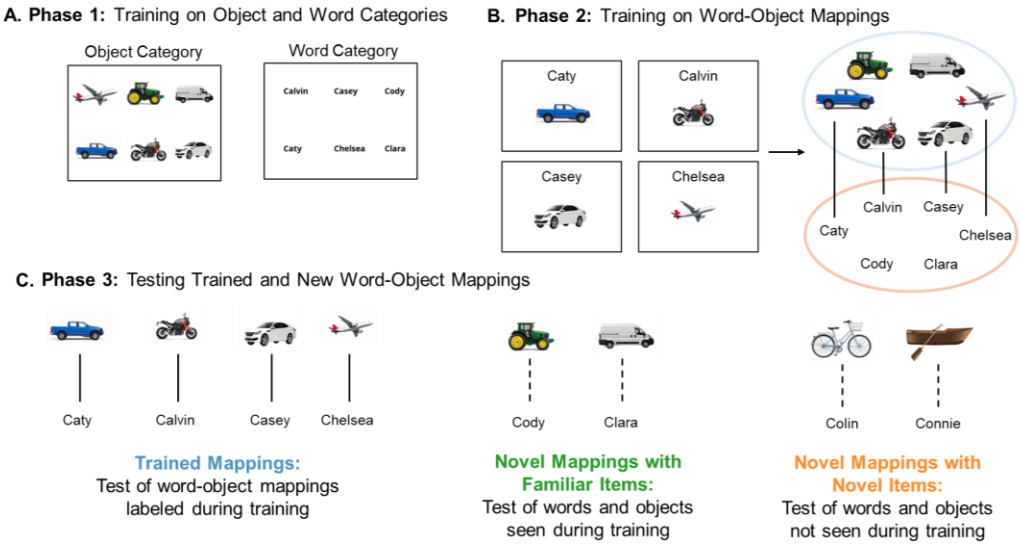

Figure 3: Procedure and example stimuli for Experiment 1. Participants learned categories of visual objects and words (A), learned the mappings between several words and objects from the two corresponding categories (B), and had to map words to objects that were trained (Trained Mappings) and that never co-occurred during training (Novel Mappings; C).

Materials

The visual stimuli included 24 images from four familiar visual categories: fruits (apple, banana, cherry, pineapple, strawberry, pear, orange, kiwi), tools (hammer, wrench, drill, broom, screwdriver, saw, tape, shovel), vehicles (car, truck, plane, motorcycle, boat, bicycle, van, and bulldozer), and clothing (shirt, pants, glove, sweater, sock, jacket, belt, scarf). These 24 images were paired with 24 disyllabic names that formed four word categories: Names with the letter "A" (e.g., Aaron, Alice), "B" (e.g., Bobby, Bella), "C" (e.g., Casey, Caty), or "D" (e.g., David, Debbie). Fruits were assigned "A" names, tools "B" names, vehicles "C" names, and clothing items "D" names. The mapping between the word and object categories (e.g., "C" names and vehicles), and the individual word-object pairs (e.g., truck - "Caty"), were assigned arbitrarily. Familiar objects and names were used instead of novel stimuli, so that learners could more quickly form the visual and word categories based on the co-occurrences during training.

Procedure

The study included three phases: Phases 1-2 (training) and Phase 3 (testing). Participants saw the same order (Figure 3).

Training (Phases 1-2) During Phase 1, participants were presented with 16 learning trials. Half of these trials were visual category trials where participants saw six objects from the same category presented together for 15 seconds (e.g., plane, bulldozer, van, truck, motorcycle, car). Half of the trials were word category trials where participants heard six names presented together for 15 seconds (e.g., Casey, Caty, Calvin, Chelsea, Cody, Clara). The object and word trials were mixed. After exposure to each visual and word category, participants were tested on their ability to group the pictures or words together via eight probe trials. These probe trials tested learning of the word and object categories, and served as an attention check. During each probe trial, participants saw six objects or six words. Half of these objects and words were from the same category, whereas the other half were from different categories. Participants were asked to select the three that went together. If they correctly grouped the items on the probe trials, they continued to Phase 2. During Phase 2, participants learned the individual mappings between words and objects. Each trial presented a single object with a single word (e.g., truck – "Caty") for 5 seconds. From among six objects in a visual category and six words in the corresponding word category, four direct word-object mappings were trained in Phase 2. Each word-object pair was trained six times, yielding a total of 96 trials (4 word-object pairs × 4 categories × 6 repetitions = 96). The remaining two pairs from training, as well as two entirely new pairs, were tested during Phase 3.

Testing (Phase 3) The testing phase comprised three test types presented in the following order: Trained Mappings, Novel Mappings with familiar items, and Novel Mappings with novel items. The Trained Mappings test trials probed participants' learning of the individual word-object mappings shown in Phase 2. On each trial, participants heard a name (e.g., "Caty") and had to select the corresponding object (e.g., truck) from among 16 objects seen during training. They were tested on each object once, resulting in 16 total trials. The Novel Mapping trials with familiar items probed participants' ability to map words and objects that they had encountered during training, but that had not been directly mapped to each other. On each trial, they heard a word (e.g., "Cody") and were asked to select the corresponding object (e.g., bulldozer) from among four objects (one from each of

the four categories). Because each presented object was from a different category, there was only one correct target object on each trial. Participants were tested on two words (out of six) for each category, yielding eight test trials. Finally, participants were presented with the Novel Mapping trials with novel items where they heard a word (e.g., "Colin") and were asked to select the corresponding object (e.g., bicycle) from among four objects (one from each of the four categories). Because each presented object was from a different category, there was only one correct target object on each trial. These words and objects had never appeared in the two training phases but belonged to the categories presented during training. Participants were again tested on two word-object mappings for each category, yielding eight test trials.

Data Analysis The study was a within-subjects design, with all participants completing all test trial types. Accuracy on each trial was coded as 1 (Correct) or 0 (Incorrect). The proportion correct for the Trained Mapping trials was computed by summing the number of correct trials and dividing by 16. The proportion correct for the Novel Mapping trials was computed by summing the number of correct trials and dividing by eight. These proportions were compared to chance level using a one-sample t-test. Linear mixed effects models in R (version 4.2.2; RStudio Team, 2019) was constructed using the *lme4* package (Bates et al., 2015) to examine whether performance on the Trained Mappings predicted performance on the Novel Mappings. The proportion correct on the Novel Mappings with familiar and novel items was regressed separately on the proportion correct for the Trained Mappings, with subject and test item category included as random slopes.

Results

First, we determined whether participants learned the direct mappings between the words and objects from training above chance level (.0625). A one-sample *t*-test revealed that participants indeed learned the words above chance on the Trained Mapping trials ($M = 0.77$, $SD = 0.42$), $t(40) = 19.54$, $p < .001$, $d = 3.01$.

Given that the direct mappings had been learned, the main question of this study was whether participants could map words to objects that they had not been trained on. As shown in Figure 4A, participants performed above chance (.25) on the Novel Mappings trials with familiar items ($M = 0.56$, $SD = 0.50$), $t(40) = 4.93$, $p < .001$, $d = 0.76$, and the Novel Mappings trials with novel items ($M = 0.64$, $SD = 0.48$), $t(40) = 7.40$, $p < .001$, $d = 1.13$. These results show that participants successfully leveraged their recently acquired knowledge of visual and word categories to infer new word-object mappings. Moreover, they did so consistently across all four categories, $ps < .001$ (Figure 4B-E).

Did participants' learning of the mappings from training influence their performance on the novel mappings? A linear mixed effects model revealed that performance on the Trained Mappings trials was not a significant predictor of performance on the Novel Mappings with familiar items, $B = 0.08$, $SE = 0.09$, $t = 0.93$, $p = 0.35$, or the Novel Mappings with novel items, $B = 0.004$, $SE = 0.01$, $t = 0.03$, $p = 0.97$. This finding suggests that the strength of participants' word-object mappings from training did not influence their success in mapping novel words and objects at test. Even if they failed to map the words and objects from training, they could use the learned correspondence between the object and word categories (e.g., "C" names go with vehicles) to successfully infer new word-referent mappings at test.

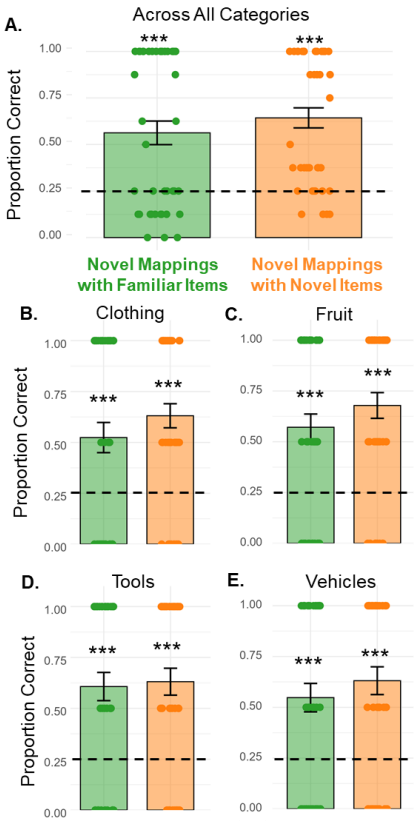

Figure 4: Average proportion correct on the Novel Mappings with familiar items and Novel Mappings with novel items (A), also visualized by category (B-E). Dashed line represents chance level. ***, $p < .001$.

Together, Experiment 1 demonstrated that adult learners could track and integrate co-occurrences between words, objects, and words and objects to learn two types of novel word-object mappings. These co-occurring items also shared higher-order category structure, which is more representative of real-world learning environments. When trained on a subset of individual word-object mappings, participants also inferred the correspondence between the word and object categories and used this accumulated knowledge to engage in zero-shot learning.

Experiment 2

This study examined whether learners could also integrate acquired co-occurrences between words, as well as words and objects, with the semantic similarity among visual objects to

build new word-object mappings without being directly trained. As shown in Figure 5, there were three experimental conditions. In the Label-Only condition, a familiar word (e.g., "cow") co-occurred with a novel word (e.g., "dalele") in training. At test, participants were presented with visual objects from different categories, including a target object (e.g., bison) that belonged to the same category as the familiar word ("cow"). If learners utilized the visual similarity between the familiar object (cow) and the target object (bison), and the co-occurrence between the familiar word ("cow") and the novel word ("dalele"), then they should correctly choose the target object (bison) over other available objects from different categories. In the Visual-Only condition, participants were presented with a familiar visual object (cow) and a novel word ("dalele"). At test, they should also choose the target object (bison) that is visually similar to the familiar object (cow) for the novel word ("dalele"). In the Visual-Label condition, both a familiar object (cow) and a familiar word ("cow") co-occurred with a novel word ("dalele") during training. Again, learners should choose the visually similar object over other available objects at test.

Participants

The participants were a new sample of 40 adults recruited from Prolific and were compensated monetarily for their participation. Participants were on average 38.4 years old (*SD* = 12.9, range: 22-71 years) and included 22 females. A majority (60%) identified as White (20%: Black or African American, 15%: declined to respond, 5%: Asian). A majority (62.5%) also reported holding a bachelor's or graduate degree (17.5%: some college, 12.5%: declined to respond, 5%: high school degree, 5%: associate's degree).

Materials

The stimuli included 24 novel words, 16 familiar words, and 40 images of familiar objects. The images and words were assigned to one of three conditions: Label-Only, Visual-Only, or Visual-Label. In each condition, there were eight training items from different basic level categories and eight word-object mappings to identify at test.

Procedure

Participants participated in each of three conditions. Each condition included a training phase followed immediately by a testing phase (Figure 5).

Label-Only Condition In this condition, participants heard a novel word paired with a familiar word on each training trial, but did not see an object. Each 7-second trial only presented a cross-hair and the two words auditorily. In total, participants heard eight word pairs (e.g. "monkey-*blicket*", "apple-*jefa*") presented 12 times for a total of 96 trials. Participants then entered the testing phase where they heard one of the eight novel words and saw eight entirely new objects (e.g., lemur, mango etc.). They were asked to select the most likely object for the novel word. One object was from the same category—and shared perceptual features

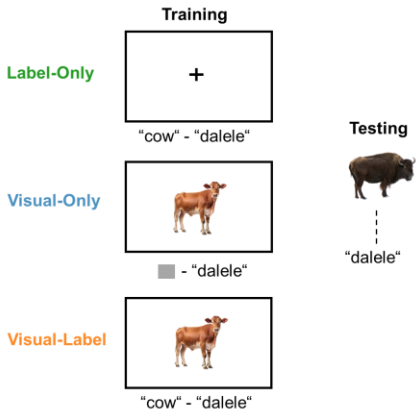

Figure 5: Illustrative example stimuli from Experiment 2. Depending on the condition, novel words were paired with familiar words, objects, or both during training. Participants had to map the novel words to a perceptual match at test

with—the familiar item that was paired with the novel word, but had not occurred with the novel word during training. Participants thus had to integrate their newly acquired knowledge of the word pairs with their prior knowledge of perceptual similarity between objects to correctly map the novel words to previously unseen objects.

Visual-Only Condition In this condition, participants heard a novel word and saw an image of a familiar object on each 7-second trial. Each of the eight novel word and object pairs (e.g., "cow-*dalele*", "house-*pribble*") was presented 12 times for a total of 96 trials. During the testing phase, participants heard one of the novel words and saw eight entirely new objects (e.g., bison, gazebo etc.). They were asked to select the most likely object for the novel word and completed one test trial for each of the eight novel words.

Visual-Label Condition The Visual-Label condition was a combination of the previous two conditions. During each 7-second trial, participants saw a familiar object and heard two words: a familiar word that corresponded to the object and a novel word. Each of the eight word pairs and their corresponding objects (e.g., "fish-*kenofe*", "bike-*modi*") were presented 12 times for a total of 96 trials. Participants then entered the testing phase where they heard one of the novel words and saw eight entirely new objects (e.g., stingray, scooter etc.). They were again asked to select the most likely object for the novel word and completed one test trial for each of the eight novel words.

Data Analysis The study was a within-subjects design, with all participants completing all conditions. Accuracy on each trial was coded as Correct (1) or Incorrect (0). Proportion correct at test for each condition was computed by summing the number of correct trials and dividing by eight test trials. These proportions were compared against chance level using one-sample t-tests. We also constructed a logistic mixed-effects model in R (version 4.2.2; RStudio Team, 2019) using

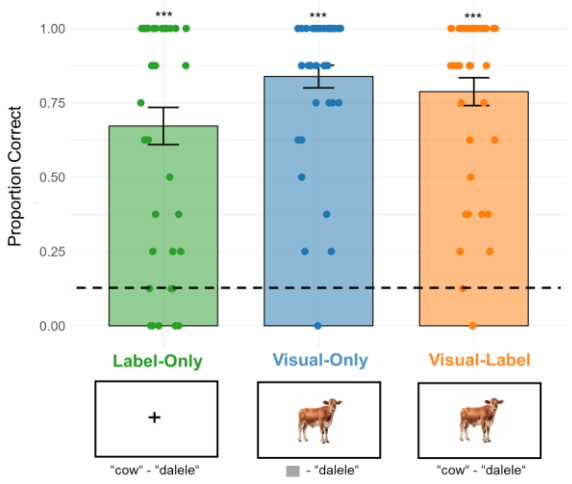

Figure 6: Proportion correct by condition (Label-Only, Visual-Only, Visual-Label). Dotted line represents chance level (.125). ***, p <.001.

the *lme4* package (Bates et al., 2015) to examine whether certain training conditions led to higher performance than others. Accuracy on each test trial was regressed on the treatment-coded condition (Label-Only served as the reference group), and subject and test item were included as random slopes.

Results

The main question of this study was whether participants could integrate visual and auditory co-occurrences from training with their prior knowledge to successfully map novel words to objects at test. One-sample t-tests confirmed that participants were indeed above chance (0.125) for the Visual-Only (M = 0.84, SD = 0.37), $t(39)$ = 18.71, p < .001, d = 2.92, Label-Only (M = 0.67, SD = 0.40), $t(39)$ = 8.73, p < .001, d = 1.38, and Visual-Label (M = 0.79, SD = 0.30), $t(39)$ = 14.15, p < .001, d = 2.24, conditions (Figure 6).

Next, we examined whether certain training conditions led to higher performance than others. A logistic mixed effects model, Accuracy ~ 1 + Condition + (1 | ID) + (1 | Test Item), revealed a main effect of Condition, such that participants were 4.74 times more likely to be correct in the Visual-Only condition than the Label condition, B = 1.56, SE = 0.30, z = 5.16, OR = 4.74, *95% CI:* 2.62-8.56, p < .001. Similarly, participants were 2.82 times more likely to be correct on the Visual-Label condition than the Label-Only condition, B = 1.04, SE = 0.29, z = 3.60, OR = 2.82, *95% CI:* 1.61-4.96, p < .001. Although participants learned above chance in all conditions, exposure to both words and objects during training improved novel word-referent mapping, highlighting how the integration of multimodal statistical relationships accelerates zero-shot learning.

Discussion

Human learners rely on SL to extract patterns in their environment. Statistical learning mechanisms like cross-situational learning have been shown to support word learning (Smith & Yu, 2008; Yu & Smith, 2007), yet rely on substantial data accumulation to detect meaningful patterns. Can cross-situational learning mechanisms also account for the rapid acceleration in children's early word learning?

The present study provides a possible solution by showing that learners can track multiple statistical relationships between words, objects, and words and objects concurrently, and can integrate these statistics to engage in zero-shot learning—that is, inferring the meaning of new words without direct statistical evidence. In Experiment 1, we demonstrated that learners used newly acquired knowledge of visual and word categories to map words to objects that never directly co-occurred during training. Experiment 2 extended these findings by showing that learners also integrated co-occurrences between words and objects with their prior semantic knowledge of visual objects, enabling them to map novel words to perceptually similar objects. Crucially, zero-shot learning was enhanced in Experiment 2 when both words and objects co-occurred during training.

These findings suggest that learners track various statistical relationships across learning situations, such as the co-occurrences among words, the co-occurrences among objects, and the co-occurrences among words and objects. This aligns with prior studies that show learners can track higher-order statistical regularities during cross-situational word learning tasks (Chen & Yu, 2017; 2022; Dautriche & Chemla, 2014; Zettersten et al., 2018), yet extends this body of work in a key way: Learners track several types of statistics *and* integrate these statistics to accelerate new learning.

We suggest that cross-situational learning can thus become more efficient as more data accumulates, even enabling zero-shot learning without direct statistical evidence. This proposal resembles recent computational models that aligned visual and linguistic information in an unsupervised manner to facilitate learning (e.g., Roads & Love, 2020). Notably, models trained on multimodal input outperformed unimodal models on various learning tasks, including zero-shot learning (e.g., Frome et al., 2013; Kiela & Clark, 2015; Lu et al., 2019; Radford et al., 2021), reinforcing the importance of integrating multimodal statistics to accelerate learning.

There are several avenues for future work. The current study used familiar words and objects, mirroring real-world learning in which learners bring prior knowledge to the learning task. Moreover, co-occurring items also tend to share category and visual similarity in the real world. Future studies should, however, test the proposed idea by training learners on novel categories based on co-occurrences alone.

Another important future direction is to test whether children—like adults—can track and integrate several statistics to engage in zero-shot learning. This would provide compelling evidence that cross-situational learning can lead to rapid word learning with limited input. Together, these studies would demonstrate that the learning environment is noisy, yet information-rich. Fast SL thus arises from using and integrating the various kinds of statistics available in the environment, driving the efficiency in human SL.

Acknowledgements

This work was supported by NICHD R01DC017925 and R01HD104624. We thank the Developmental Intelligence Lab for their feedback on the experimental studies.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116,* 463–498.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Bolker, M. B. (2015). Package 'lme4'. C*onvergence*, *12*, 2.

Chen, C., & Yu, C. (2017). Grounding statistical learning in context: The effects of learning and retrieval contexts on cross-situational word learning. *Psychonomic Bulletin & Review, 24,* 920–926.

Chen, C., & Yu, C. (2022). Building lexical networks: Preschoolers extract different types of information in cross-situational learning. *Journal of Experimental Child Psychology, 220,* Article 105430.

Chen, C., Zhang, Y., & Yu, C. (2017). Learning object names at different hierarchical levels using cross-situational statistics. *Cognitive Science, 42,* 591–605.

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*, 468-481.

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 892.

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66-108.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499–504.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances In Neural Information Processing Systems*, *26*.

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, *17*(1), 171-183.

Kiela, D., & Clark, S. (2015). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 2461–2470

Knabe, M. L., & Vlach, H. A. (2023). Not all is forgotten: Children's associative matrices for features of a word learning episode. *Developmental Science*, *26*, e13291.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances In Neural Information Processing Systems*, *32*.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

Peters, R., & Borovsky, A. (2019). Modeling early lexico-semantic network development: Perceptual features matter most. *Journal of Experimental Psychology: General, 148(4),* 763–782.

Plunkett, K., Delle Luche, C., Hills, T., & Floccia, C. (2022). Tracking the associative boost in infancy. *Infancy*, *27*(6), 1179-1196.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning, 139.*

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence, 2,* 76–82.

Roembke, T. C., Simonetti, M. E., Koch, I., & Philipp, A. M. (2023). What have we learned from 15 years of research on cross-situational word learning? A focused review. *Frontiers in Psychology, 14*, 1175272.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 906-914.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*, 12663-12668.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52-61.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*, 52-63.

Savic, O., Unger, L., & Sloutsky, V. M. (2023). Exposure to co-occurrence regularities in language drives semantic integration of new words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48(7),* 1064–1081.

Schapiro, A., & Turk-Browne, N. (2015). Statistical learning. *Brain Mapping*, *3*, 501-506.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558-1568.

Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., & Lo, T. (2019). Routine language: Speech directed to infants during home activities. *Child Development*, *90*, 2135-2152.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional

probability. *Language Learning and Development*, *3*, 1-42.

Unger, L., Savic, O., & Sloutsky, V. M. (2020). Statistical regularities shape semantic organization throughout development. *Cognition, 198,* 104190.

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1-37.

Willits, J. A., Wojcik, E. H., Seidenberg, M. S., & Saffran, J. R. (2013). Toddlers activate lexical semantic knowledge in the absence of visual referents: Evidence from auditory priming. *Infancy*, *18*, 1053-1075.

Wojcik, E. H., & Saffran, J. R. (2013). The ontogeny of lexical networks: Toddlers encode the relationships among referents when learning novel words. *Psychological Science*, *24*, 1898-1905.

Wojcik, E. H., & Saffran, J. R. (2015). Toddlers encode similarities among novel words from meaningful sentences. *Cognition*, *138*, 10-20.

Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition* (pp. 4582-4591).

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, *4*, 32-62.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414-420.

Zettersten, M., Wojcik, E., Benitez, V. L., & Saffran, J. (2018). The company objects keep: Linking referents together during cross-situational word learning. *Journal of Memory and Language, 99,* 62–73.

Zhang, Y., Chen, C., & Yu, C. (2019). Mechanisms of cross-situational learning: Behavioral and computational evidence. *Advances in Child Development and Behavior, 56,* 37–60.