COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13305

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13305

Finding Structure in One Child's Linguistic Experience

Wentao Wang,[a] Wai Keen Vong,[a] Najoung Kim,[a,c] Brenden M. Lake[a,b]

[a]*Center for Data Science, New York University*
[b]*Department of Psychology, New York University*
[c]*Department of Linguistics, Boston University*

Received 5 December 2022; received in revised form 16 April 2023; accepted 22 May 2023

Abstract

Neural network models have recently made striking progress in natural language processing, but they are typically trained on orders of magnitude more language input than children receive. What can these neural networks, which are primarily distributional learners, learn from a naturalistic subset of a single child's experience? We examine this question using a recent longitudinal dataset collected from a single child, consisting of egocentric visual data paired with text transcripts. We train both language-only and vision-and-language neural networks and analyze the linguistic knowledge they acquire. In parallel with findings from Jeffrey Elman's seminal work, the neural networks form emergent clusters of words corresponding to syntactic (nouns, transitive and intransitive verbs) and semantic categories (e.g., animals and clothing), based solely on one child's linguistic input. The networks also acquire sensitivity to acceptability contrasts from linguistic phenomena, such as determiner-noun agreement and argument structure. We find that incorporating visual information produces an incremental gain in predicting words in context, especially for syntactic categories that are comparatively more easily grounded, such as nouns and verbs, but the underlying linguistic representations are not fundamentally altered. Our findings demonstrate which kinds of linguistic knowledge are learnable from a snapshot of a single child's real developmental experience.

Keywords: Learnability; Child development; Language learning; Statistical learning; Multimodal learning; Neural networks; First-person video

1. Introduction

In the first 3 years of life, children's linguistic development progresses rapidly. Young children begin understanding words at around 6 months (Bergelson & Swingley, 2012, 2015;

Correspondence should be sent to Wentao Wang, Center for Data Science, New York University. E-mail: wentao.wang@nyu.edu

Tincoff & Jusczyk, 1999, 2012). The vocabulary that they can comprehend and produce increases gradually until around 12–14 months, at which a nonlinear comprehension boost occurs (Bergelson, 2020) and lexical-semantic networks begin to develop (Wojcik, 2018). Language learning remains both a scientific and engineering puzzle; it is unclear what inductive biases and cognitive abilities are necessary and how much can be learned through relatively generic learning mechanisms, such as distributional learning from patterns of word co-occurrence (Firth, 1957; Harris, 1954; Landauer & Dumais, 1997).

To provide some insight into this learning challenge, we captured a subset of the linguistic and visual inputs received by a single child during their development. We then train generic computational models for sequence processing on this data and evaluate what these models learn (e.g., Orhan, Gupta, & Lake, 2020). Previously, a major obstacle to this approach was the lack of high-quality and substantive developmental data. However, thanks to large-scale developmental datasets containing linguistic input (MacWhinney, 2000; Roy, Frank, DeCamp, Miller, & Roy, 2015; Sullivan, Mei, Perfors, Wojcik, & Frank, 2021) and recent advances in deep learning, it is now possible to run large-scale simulations on real language input. Training neural networks on these datasets, and then analyzing what kinds of knowledge are acquired, can help to answer foundational questions about what aspects of language are learnable from a child's experience (Huebner & Willits, 2018; Warstadt & Bowman, 2022) via primarily distributional learning, without social cognition abilities and aspects of world knowledge that are thought to play central roles (Bloom, 2000; Markman, 1989; Murphy, 2002).

In this work, we follow this approach by using SAYCam, a recent longitudinal developmental dataset consisting of an egocentric visual and linguistic input to a single child spanning 6–25 months of age (Sullivan et al., 2021). The scale of this dataset allows us to train several widely used neural network architectures and explore what they learn, in terms of how they structure their representations and how this affects behavior. The networks we adopt are not designed for human languages specifically; rather, they are configured to process general sequences. We first train two kinds of neural networks, Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber, 1997) and Continuous Bag-Of-Words (CBOW; Mikolov, Chen, Corrado, & Dean, 2013), on only the language portion of the dataset and analyze the syntactic and semantic structure they acquire. Then, we add the visual data and train an image captioning model (Xu et al., 2015) on the paired vision-and-language dataset, and examine the impact on linguistic knowledge from incorporating the visual modality.

Our work builds on previous examinations of what computational models can learn from linguistic input (Abend, Kwiatkowski, Smith, Goldwater, & Steedman, 2017; Elman, 1990; Huebner & Willits, 2018; Huebner, Sulem, Cynthia, & Roth, 2021; Perfors, Tenenbaum, & Regier, 2011, *i.a.*). In his pioneering article, Elman (1990) formulated a means of training simple recurrent networks (SRNs) to predict the next word in a sentence given the previous words. When applied to simple language-like inputs, these networks formed coherent clusters of words, analogous to real English syntactic and semantic categories. More recently, researchers have examined similar questions using naturalistic sources of data combined with more capable neural network architectures, such as LSTMs (Hochreiter & Schmidhuber, 1997) and Transformers (Vaswani et al., 2017). For instance, Huebner and Willits (2018)

trained both Elman's SRNs and LSTMs on a corpus of naturalistic, developmental linguistic data (CHILDES; MacWhinney, 2000), and analyzed emergent clusters in their acquired representations. Similarly, Huebner et al. (2021) trained a Transformer on a corpus derived from CHILDES (AO-CHILDES; Huebner & Willits, 2021) and analyzed its syntactic knowledge. Other related work has focused on learning structured probabilistic models from naturalistic linguistic inputs, using methods based on probabilistic grammar induction to learn syntactic structure and word meanings (Abend et al., 2017; Perfors et al., 2011; Waterfall, Sandbank, Onnis, & Edelman, 2010). Our work follows in these modeling traditions, as exemplified by Elman's seminal work. The most distinctive aspect of our work is that the networks are trained on a strict subset of real developmental experience from just one child, without using outside annotation beyond the transcripts. Previous work in this vein either aggregated linguistic input across multiple children or utilized structured representations and/or annotations to help bootstrap learning. Thus, our work provides a unique window into the learnability of linguistic structure based on one child's input—without additional data and labels, using a distributional learning strategy.

We conduct a range of linguistic evaluations and find our models achieve varying degrees of success across different linguistic phenomena. When using language-only data, we find that networks can differentiate words in different syntactic categories, such as nouns, transitive and intransitive verbs, and semantic categories, such as animals and clothing.[1] We also find that these networks acquire nascent syntactic abilities, such as inferring the syntactic category of a word from its context. In some cases, they can recognize determiner-noun agreement and argument structure regarding verb transitivity, but they struggle with other phenomena, such as subject-verb agreement. Additionally, we find that introducing visual information provides an incremental improvement on our networks' abilities to predict words in context, but does not fundamentally alter the linguistic representations.[2]

2. Sensory input through the eyes and ears of a child

In this section, we briefly describe the data streams used for training and evaluating our neural networks. The data are a subset of SAYCam (Sullivan et al., 2021), a dataset consisting of egocentric head-mounted camera recordings of three very young, English-speaking children.[3] Each child's recordings are recorded at regular intervals (several hours each week) for around 2 years starting from 6 to 8 months of age. However, out of the three children, only one (labeled as baby S) had a large proportion of his naturalistic speech input transcribed (spanning 6–25 months of age), making baby S the choice of our focus. This dataset, which we call the **SAYCam-S** dataset, consists of child-directed utterances paired with visual data from the child's point of view at the time of the utterance.

We outline the major steps taken to preprocess the dataset. For each original transcript, we first replace anything annotated as "inaudible" with a special <UNK> (unknown) token, and use the spaCy tokenizer (Honnibal & Montani, 2017) to segment the inputs into discrete tokens. Moreover, long utterances were split into multiple sentences, and their time spans were obtained by linearly interpolating the time span of the original transcript.[4] We filter the

Table 1
Statistics of SAYCam-S

	Train	Validation	Test
Number of utterances	33,737	1874	1875
Mean (SD) utterance length	6.67 (5.49)	6.59 (5.46)	6.62 (4.95)
Number of tokens	225,001	12,355	12,418
Number of frames	540,681	29,686	29,918
Mean frames per utterance	16.0	15.8	16.0
Out-of-vocabulary rate	1.99%	2.42%	2.79%

utterances by excluding child-produced utterances, retaining only those from parents to focus on the input that the child receives. For each utterance, we extract multiple frames at 5 frames per second (fps) from the video, up to the first 6.4 s of its time span.

The dataset is randomly split into training, validation, and test sets (90%/5%/5% of all utterances, respectively).[5] In this study, only the training and validation sets are used, while the test set is left for future use. Our vocabulary is built from all tokens contained in the training set, excluding those with a frequency less than 3 in this set, resulting in a final vocabulary size of 2350. Any out-of-vocabulary tokens are replaced by the special <UNK> token. Appendix A.1 contains additional details.

The preprocessed dataset consists of 37,486 child-directed utterances (249,774 tokens) paired with 600,285 image frames. Table 1 contains further descriptive statistics about the dataset, and Fig. 1 shows some sample frames from the dataset paired with their corresponding utterances. Notably, the average utterance length is rather short compared to sentence lengths in typical written corpora, which is a characteristic of child-directed speech.

3. Neural networks and training

In this section, we introduce three kinds of neural network architectures and how we train them on our SAYCam-S dataset. Note that because we are studying what is learnable in principle from one child's linguistic experience, we do not constrain ourselves to network architectures and training configurations that are strictly biologically or psychologically plausible. One reason is that these questions are still open: we are far from a mature understanding of the algorithmic issues involved in modeling individual cognitive development from realistic input over the timescales of years (including the contributions of multiple memory systems, constraints of attention, etc.). Instead, we use common machine learning architectures and training practices that are known to be effective, leaving the integration of cognitive constraints as an avenue for future work.

3.1. Language-only networks

We use two kinds of networks to encode the language input: single-layer uni-directional LSTM (Hochreiter & Schmidhuber, 1997), which is a variant of recurrent neural network

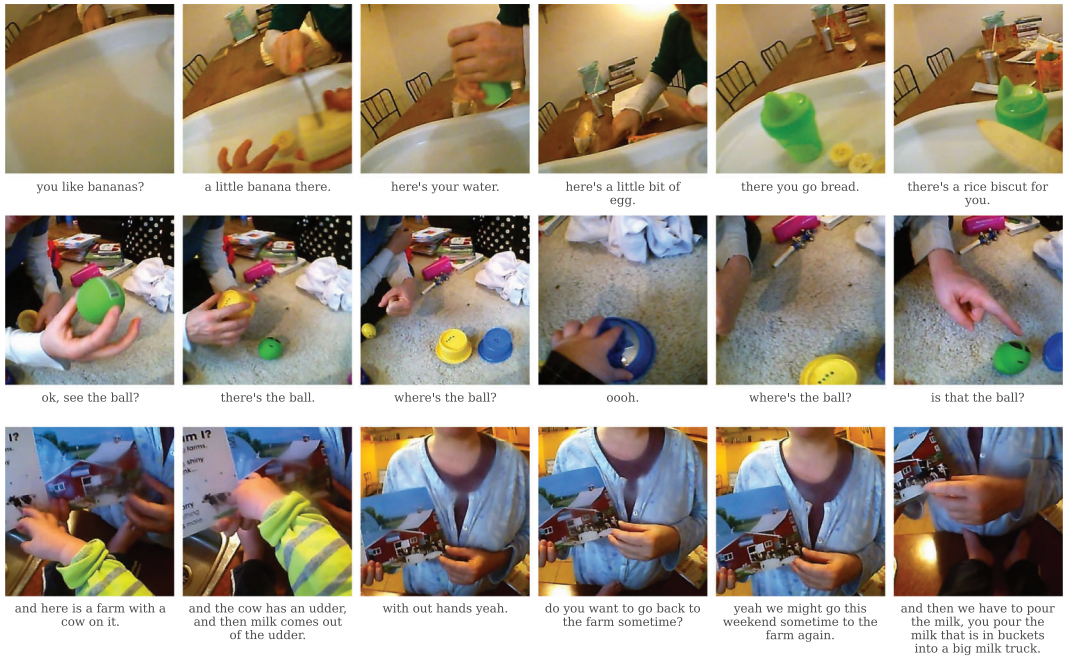

Fig. 1. Example frames and their corresponding utterances. Each row is a different scene: eating breakfast, playing a game with a ball, and reading a farm-themed picture book. Unlike common image-text datasets in machine learning, the utterances only loosely align to the frames. For instance, the foods mentioned in the utterance are not always in the corresponding video frames, and the ball mentioned in the utterance is sometimes covered by the cup.

(RNN), and CBOW (Mikolov et al., 2013). The neural networks are trained from scratch: their training objective is token prediction in context using a cross-entropy loss, which involves multiple sweeps through the dataset during the training process.

Fig. 2(b) illustrates the architecture of a uni-directional LSTM. A uni-directional LSTM processes a sequence of tokens left-to-right, and maintains a hidden state after each step, keeping track of context using only tokens to the left of the predicted token in the utterance. The dimensions of the hidden states and the word embeddings are both 512.[6] When predicting the next token, the LSTM assigns a probability distribution over all tokens in the vocabulary.

Fig. 2(a) illustrates the CBOW architecture. For CBOW, the context it can see is a constant number of tokens to the left and right of the predicted token. The set of these tokens is called its "context window." One advantage this provides over uni-directional networks is that the CBOW can additionally utilize information from the right of the token to be predicted. However, unlike the LSTM, its context window size is fixed to a small number, preventing it from modeling long-distance dependencies. CBOW also has a simpler architecture compared to the LSTM: it uses an embedding layer to first embed the discrete input tokens into their word embeddings. Then, all word embeddings within the context window are averaged and then projected by an output layer, producing the predicted distribution over all tokens.

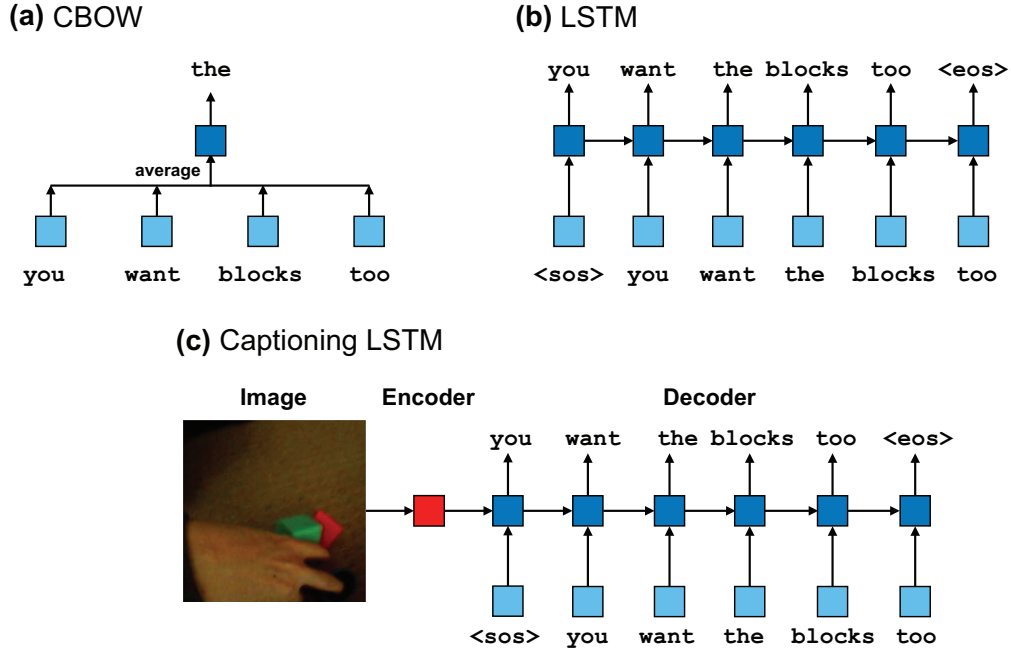

Fig. 2. The three neural network architectures. (a) The CBOW network predicts a missing word given a surrounding context of fixed size. The LSTM (b) and Captioning LSTM (c) networks both predict the next word given a sequence of previous words (additionally a corresponding image for the Captioning LSTM). The light blue boxes indicate word embeddings, the dark blue boxes indicate hidden embeddings, and the red box indicates the visual embedding. Fig. adapted from Lake and Murphy (2021).

All embeddings are of size 512. All parameters of both the LSTM and the CBOW, including the input and output embeddings, are randomly initialized. See Appendix A.2 for additional details regarding network architectures and training configurations.

We measure these networks' performance on token prediction by per-token perplexity.[7] Our LSTM and CBOW models reached an average perplexity of 24.80 ($SD = 0.21$) and 22.20 ($SD = 0.01$) on the validation set, respectively, averaged over three runs with different random seeds.[8] Despite the benefit of incorporating bidirectional context, CBOW is only marginally better than the LSTM on this measure. For CBOW, we tested context window sizes ranging between 1 and 4 tokens on both sides of the predicted token and found that a context window containing only one token on both sides performed best.[9]

3.2. Multimodal network

Another advantage of SAYCam-S is its multimodality: it contains parallel vision and language inputs. Adding visual information provides grounding for words, potentially allowing the networks to learn references from words to objects, or at least visual features in the input (Hill et al., 2021; Vong & Lake, 2022). Multimodal learning has been shown to help resolve ambiguities when only linguistic information is present (Berzak, Barbu, Harari, Katz, &

Ullman, 2015; Christie et al., 2016), induce constituent structures (Shi, Mao, Gimpel, & Livescu, 2019), and ground events described in language to video (Siddharth, Barbu, & Siskind, 2014; Yu, Siddharth, Barbu, & Siskind, 2015).

As a way to incorporate the aligned visual modality for in-context token prediction, we treat each utterance as the caption of its associated frames. We then build an image captioning network (Xu et al., 2015), which is a uni-directional LSTM with the same architecture as described above, with an additional capacity to process information from visual inputs. This Captioning LSTM architecture is illustrated in Fig. 2(c). We use a convolutional neural network as our vision encoder (specifically, ResNeXt-50 32x4d; Xie, Girshick, Dollár, Tu, & He, 2017), pretrained via unsupervised learning from the visual stream of child S (the single child we focus on) in SAYCam (Orhan et al., 2020). The visual representation produced by the vision encoder is used to initialize the hidden state of the uni-directional LSTM. Compared to the text-only LSTM, the captioning network shares the same LSTM architecture for language processing and is trained to optimize the same objective, next token prediction. Therefore, it provides a natural comparison: we can apply the same set of linguistic analyses to both models and potentially isolate the contribution of multimodality. See Appendix A.2 for additional details.

The perplexity of our Captioning LSTM was 22.10 ($SD = 0.20$) averaged over three runs, which was incrementally lower than the language-only LSTM, suggesting a minor benefit of information from the additional visual modality. Noise in the alignment between the visual and language streams likely damped the size of the improvement. We discuss this issue further in the context of the limitations of the multimodal objective in the General Discussion.

4. Results

4.1. Learning from language only

4.1.1. Syntactic and semantic categories

Our initial analyses closely follow Elman (1990)'s approach to assessing emergent linguistic structure in neural networks. Thus, before discussing our results, we briefly summarize what Elman found. Elman trained SRNs on synthetic language data and then fit cluster dendrograms to the hidden layer activation patterns. Elman demonstrated the emergence of soft, hierarchical category structures of words: two large categories for nouns and verbs, and finer subcategories for each of them, including animate vs. inanimate nouns and transitive vs. intransitive verbs.

In our results, we find that neural networks trained on SAYCam-S show similar emergent syntactic and semantic category structures. We demonstrate this in three separate analyses, visualizing the plots of the first two analyses on the LSTM in the main text and the corresponding plots for CBOW (and Captioning LSTM) can be found in Appendix A.4. First, as in Elman (1990)'s SRN, we find that representations learned by the LSTM and CBOW form clusters corresponding to syntactic categories, including nouns and verbs. The verbs also form finer subcategories, including transitive and intransitive verbs. These findings are shown in Fig. 3; we visualize the LSTM's word embeddings using t-SNE (van der Maaten &

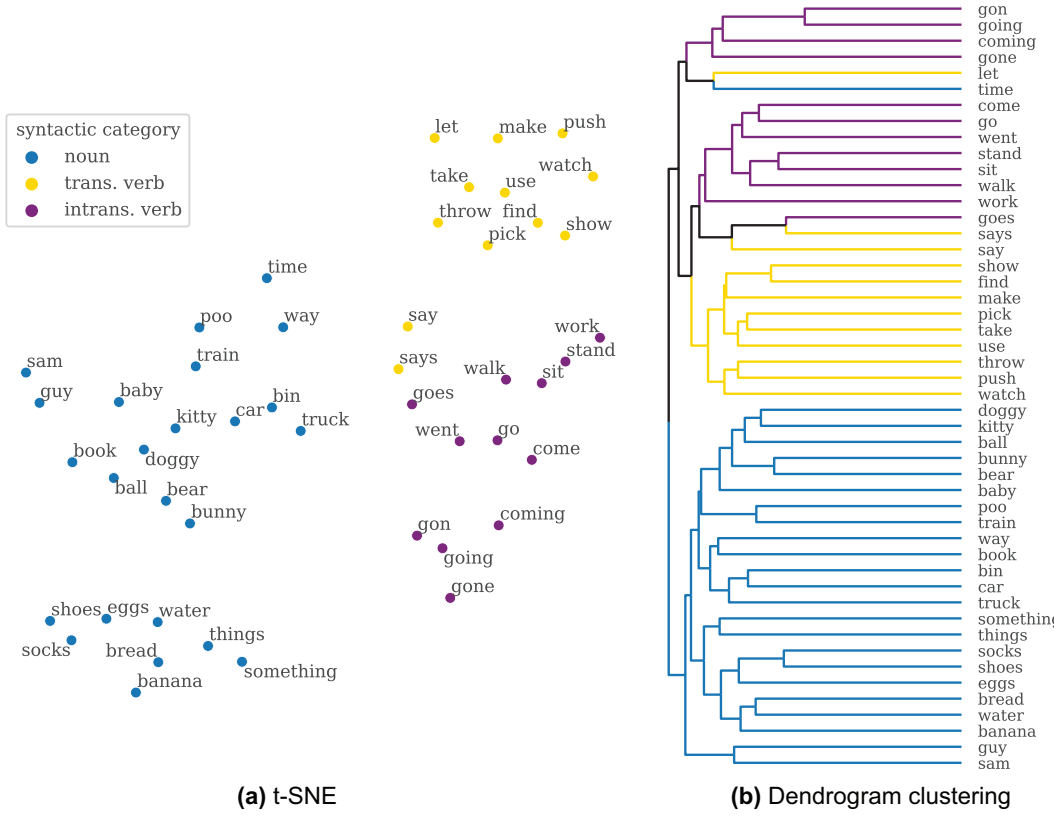

Fig. 3. Clustering LSTM's word embeddings for syntactic categories. For two embeddings u, v, t-SNE uses $1 - \cos(u, v)$ as the distance metric, and dendrogram uses $\cos(u, v)$ as the similarity measure. Nouns and verbs form two large clusters. Transitive and intransitive verbs form two smaller subclusters.

Hinton, 2008) and a dendrogram for the most frequent 24 nouns—12 transitive verbs and 12 intransitive verbs that are unambiguous in their transitivity[10] (see Fig. 8 in the Appendix for CBOW results). Both the t-SNE and dendrogram use cosine-based metrics between word embeddings.[11] Furthermore, Figures 12 and 13 in Appendix demonstrate that clusters for other syntactic categories like adjectives and adverbs also emerge from training. Interestingly, although CBOW is much simpler than the LSTM, its emergent syntactic clusters are just as clear.

Second, we find that the representations learned by the LSTM form clusters corresponding to semantic subcategories of nouns. We manually label the most frequent nouns that are unambiguously in different semantic categories, using a reference set of semantic categories derived from WordBank (Frank, Braginsky, Yurovsky, & Marchman, 2016).[12] We exclude categories having less than six unambiguous words from our analysis. As can be seen from Fig. 4, there are several visually identifiable clusters that correspond to different semantic categories.[13] In addition, we find evidence for an animate vs. inanimate distinction among this set of nouns (Elman, 1990), but this distinction is closely aligned with the semantic

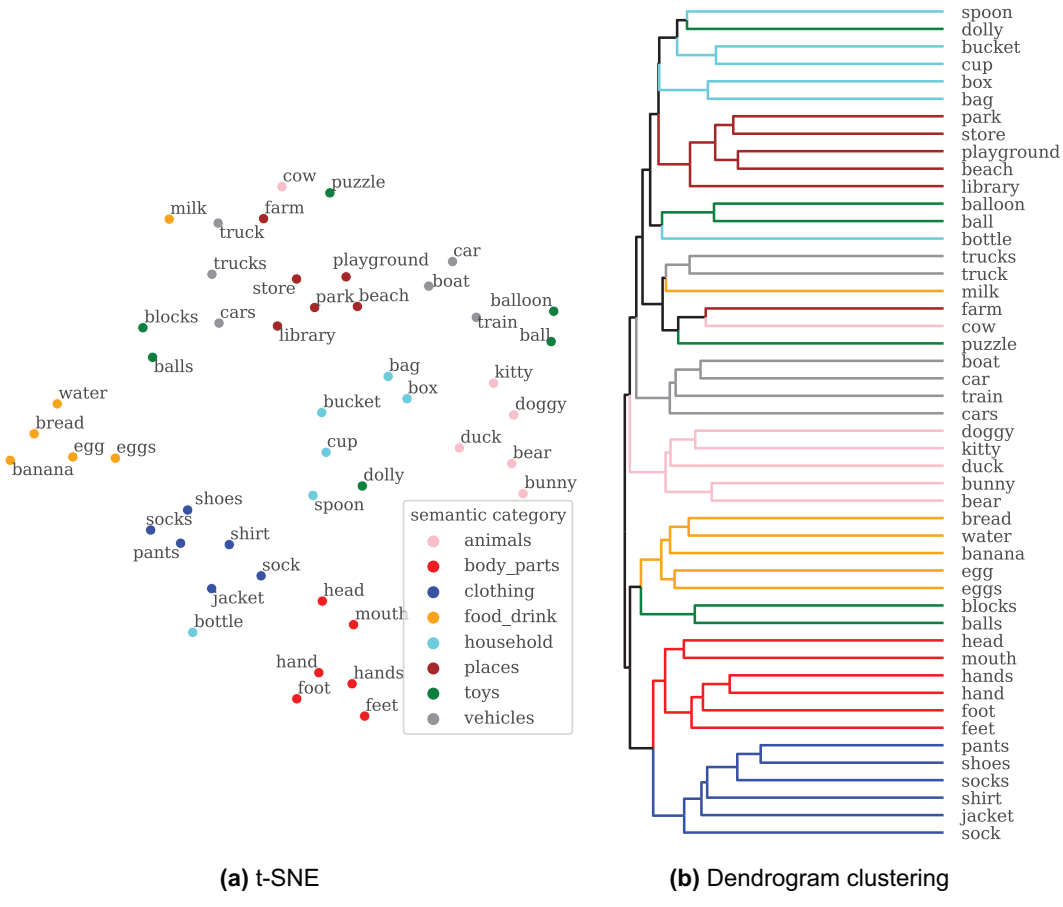

Fig. 4. Clustering LSTM's word embeddings for semantic categories. Again, both plots use cosine measures in Fig. 3. We present the most frequent six words from eight different categories. Most distinct clusters clearly correspond to semantic categories.

category structures (see Figure 15 in the Appendix). Interestingly, some thematically related words ("milk," "farm," and "cow") are close to each other. We find that this cluster can be directly traced back to a particular book in the training data; these words co-occur in scenes where the parent is reading a farm-themed picture book, illustrated in the third row of Fig. 1.

Third, as pointed out by Linzen and Baroni (2021), information in the representation may not be used by the network to causally affect its behavior. We, therefore, apply additional behavioral tests to provide further evidence for syntactic category structures in our networks. We design a novel cloze test (Taylor, 1953) to evaluate the noun-verb distinction. We build clozes such as "we are going to ___ here," where the cloze expects either a noun or a verb.[14] Trials are generated by iterating over utterances in the validation set, identifying each token that is a noun or verb, and replacing one of these tokens with an empty slot to create a cloze. For each cloze, we fill the slot with every possible noun or verb in the vocabulary, scoring each candidate with the whole-sequence probability. After normalizing these scores such that

Table 2
Examples of clozes and the networks' predictions

Model	Top-5 predictions									
we should turn on some lights, huh?										
LSTM	91.2%	put	5.2%	turn	0.4%	leave	0.4%	keep	0.4%	get
CBOW	48.2%	put	31.4%	lid	8.9%	go	2.3%	sit	1.9%	come
we should turn on some lights, huh?										
LSTM	14.0%	lights	13.4%	toys	9.5%	water	7.6%	music	5.4%	books
CBOW	11.3%	ducks	10.2%	bread	8.0%	breaky	5.8%	books	5.1%	grapes
are you done going potty?										
LSTM	9.3%	done	6.4%	're	6.0%	feeling	5.5%	hiding	5.4%	are
CBOW	69.1%	're	26.4%	re	4.2%	are	0.1%	keep	0.1%	were
and there's a kitty looking at a mouse.										
LSTM	40.9%	kitty	18.9%	mouse	4.3%	doggy	3.8%	door	2.3%	dog
CBOW	23.0%	lot	4.9%	bit	3.5%	bottle	3.0%	tower	3.0%	banana
we might go to the beach today.										
LSTM	61.2%	library	10.1%	playground	8.8%	beach	2.9%	park	2.9%	farm
CBOW	37.0%	library	22.3%	beach	17.3%	camera	12.7%	garden	4.0%	farm
now on our way we can get some food for us for breakfast										
LSTM	56.2%	bread	6.9%	chicken	4.2%	strawberries	4.0%	water	3.9%	salmon
CBOW	12.6%	lunch	11.6%	breaky	11.4%	dinner	6.9%	oil	6.0%	clothes

Note. We present a cloze by underlining the ground-truth word at the slot. We list the top-5 predictions in this form: (predicted normalized probability, word). The top predictions frequently align with expected categories. For instance, a noun follows a determiner, and a word in the food-drink category occurs if breakfast is mentioned. By comparing the predictions of the LSTM and the CBOW, we can also see the disadvantages of CBOW's small context window. For instance, in the fourth example, the CBOW model could not see the word "kitty" farther away, so it could not make a more reasonable guess that the word at the slot should be in the animal category as the LSTM did.

they sum to 1, we can estimate the degree to which the network anticipates a noun or verb in a particular slot. Across the 2412 clozes we generated (with a base rate of 65% verbs), LSTM achieves a high accuracy of 97.96% ($SD = 0.23\%$ over three runs) and CBOW achieves an accuracy of 91.20% ($SD = 0.33\%$). Table 2 presents some cloze examples and top predictions from our networks. Appendix A.5 contains more details regarding cloze construction and additional examples. Overall, these results demonstrate the network's ability to contextually differentiate nouns and verbs, supplementing our earlier findings.

4.1.2. Linguistic acceptability analysis

Next, we examine the networks' sensitivity to acceptability of a sequence modulated by more complex linguistic phenomena, such as subject-verb agreement and argument structure, again following Elman's lead (1989, 1991). We study this using Zorro: a minimal pair test

suite for 13 different linguistic phenomena (Zorro; Huebner et al., 2021), which itself is derived from another minimal pair test suite (BLiMP; Warstadt et al., 2020). The minimal pair approach asks models to judge which of the two sentences is more acceptable (e.g., "I saw this toy" vs. "I saw this toys"). The sentences in a minimal pair highlight a single linguistic phenomenon that leads to a contrast in acceptability judgments. We filter the Zorro dataset such that only sentence pairs that are entirely within our models' vocabulary are included. This leaves us with 15 subsets of the dataset, corresponding to seven different linguistic phenomena; eight were excluded for having no items after filtering. Additional details regarding dataset curation can be found in Appendix A.6.

On these filtered subsets, we test and compare several networks: the three networks we trained (language-only LSTM, CBOW, and Captioning LSTM[15]), two baseline *n*-gram language models based on statistics of the training set (unigram and bigram language models[16]), and a strong Transformer model (pretrained weights from Huebner et al., 2021 trained on AO-CHILDES which aggregates data from many children). The results are summarized in Fig. 5. Though the networks trained on SAYCam-S perform worse than the Transformer trained on more data, they are clearly above chance on many tests. For example, the LSTM achieves 67.7% accuracy on determiner-noun agreement, and the CBOW achieves 61.1% accuracy. The lower performance of CBOW on this test can be explained by the length of the dependency that needs to be processed. That is, some of the dependencies in this test span longer distances than CBOW's context window, which is advantageous for the LSTM. However, on the subject-verb agreement test which requires even longer dependencies, even the LSTM does not perform substantially above chance (55.7%). It is possible that there are too few distributional cues for long-distance agreements in SAYCam-S in particular; other findings have also shown that RNNs (Elman, 1991; Linzen & Leonard, 2018) and Transformers (Pérez-Mayos, Ballesteros, & Wanner, 2021; Tay et al., 2021) with modest amounts of training data in general have difficulty with longer-distance dependencies.[17] Other tests such as quantifiers and grammatical case are less useful for distinguishing between models because the unigram and bigram models performed well, indicating that even very simple distributional statistics are sufficient for high accuracy on these tests. See Appendix A.6 for a more detailed explanation of baseline *n*-gram models and further analysis of the relative performance of different models.

4.2. Learning from multimodal input

As mentioned earlier, the LSTM showed an incremental improvement in perplexity with additional visual information. In this final set of analyses, we examine how incorporating visual information influences the linguistic representations in the Captioning LSTM.

4.2.1. Sources of multimodal improvement

To investigate the areas of possible improvement, we first measure the improvement in cross-entropy loss for words occurring at least twice in the validation set, grouped by each word's syntactic category. This difference in loss between the Captioning LSTM and the language-only LSTM is shown in Fig. 6. The improvements for most syntactic categories are

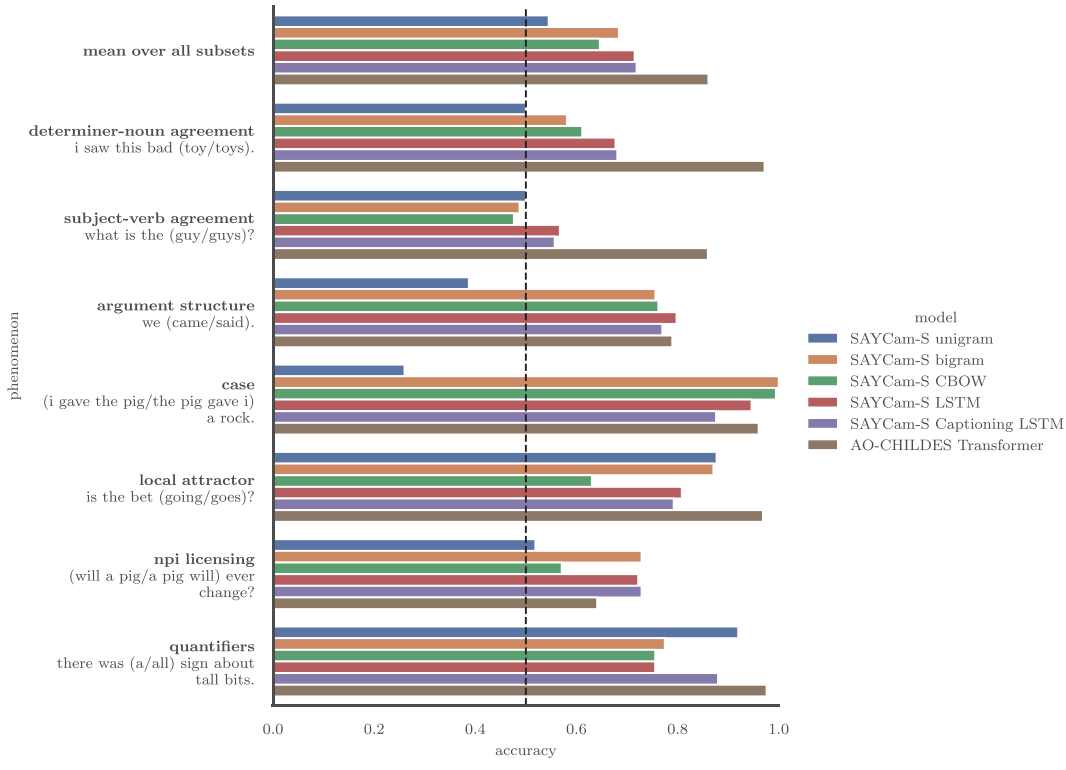

Fig. 5. Mean linguistic acceptability test accuracy over subsets for each network and linguistic phenomenon. The top group of bars is the mean over all subsets, and each of the remaining groups is the mean over the subsets corresponding to specific linguistic phenomena. Each label for a phenomenon is accompanied by an illustrative example, in which the first option in the bracket is grammatical, while the second is not. The model is correct if it assigns a higher probability to the grammatical sentence over the ungrammatical one. The dashed line denotes chance accuracy. See Appendix A.6 for fine-grained results on each phenomenon.

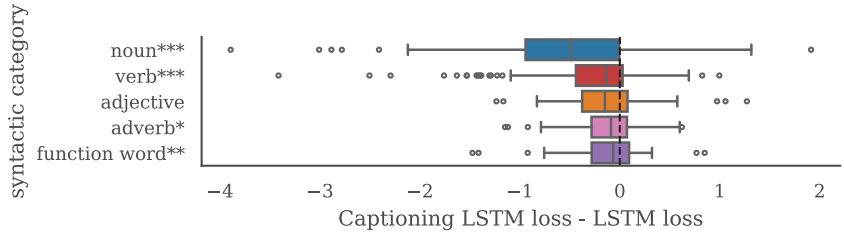

Fig. 6. Type-level loss difference between language-only LSTMs and Captioning LSTMs on the validation set. Losses are means over all occurrences of the word type and all three runs for each architecture. The box plot shows the median, first quartile, and third quartile; the whiskers extend from the box by 1.5x the inter-quartile range. *: $p \leq .05$, **: $p \leq .01$, ***: $p \leq .001$. More negative values on the x-axis indicate more improvement with added visual information. See Table 7 in the Appendix for detailed *t*-test results.

statistically significant (Table 7 in the Appendix), but in particular, nouns and verbs benefit the most from additional visual information. The improvement for nouns is expected, since most nouns acquired early by children can be visually grounded (Frank, Braginsky, Yurovsky, & Marchman, 2021). Surprisingly, verbs and even function words show some improvement, even though they are often more challenging to directly ground in images.

It is challenging to discern precisely which visual-linguistic correlations are responsible for the improved predictive power. Nevertheless, in Fig. 7, we provide several examples and compare the cross-entropy losses of the text-only LSTM and Captioning LSTM on each token of the utterances. For concrete nouns like "ball" in the third example, introducing frames containing clear referents greatly reduces losses on them. In other examples, however, the influence of visual information is not clearly beneficial or interpretable. For example, in the fourth example, the loss on "car" decreased, but the loss on "ball" increased despite both referents being present in the frame. This suggests the network also acquires less interpretable and indirect visual-linguistic correlations. One possible hypothesis for the additional improvements in cases where there are no direct referents in the scene is that different visual moments in childhood (e.g., mealtime vs. play) elicit sufficiently different distributions of words (Roy et al., 2015). The seventh example is an illustration of such a case. We leave further investigation in this direction for future work.

4.2.2. Influence on representations

As a second analysis on how visual information influences linguistic representations, we perform Representational Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008) across the three neural networks. We compute the dissimilarity matrices of the three networks' representations for the set of words in the aforementioned syntactic category analysis in Section 4.1.1, using a dissimilarity metric: $\frac{1}{2}(1 - \cos(u, v))$. Visualizations of these matrices can be found in Appendix A.8.

The similarity between representations of two networks is the Pearson correlation between elements in the upper triangulars of their dissimilarity matrices. The two networks based on the same LSTM architecture (language-only LSTM and Captioning LSTM) are quite similar to each other ($r(1126) = .82,\ p < .001$), while CBOW is less similar to either LSTM ($r(1126) = .71,\ p < .001$ to LSTM, $r(1126) = .70,\ p < .001$ to Captioning LSTM). The high similarity between the LSTM and Captioning LSTM is consistent with recent studies which found that incorporating visual information does not dramatically restructure or improve linguistic representations (Iki & Aizawa, 2021; Yun, Sun, & Pavlick, 2021).

5. General discussion

Our work demonstrates what kinds of linguistic knowledge are learnable from the naturalistic input received by a single child. There are three main takeaways. First, using the SAYCam dataset (Sullivan et al., 2021) and techniques from modern machine learning and natural language processing (NLP), we find that neural networks learning exclusively from developmentally plausible data can differentiate words in different syntactic categories. These categories

Fig. 7. Predicting an utterance with (Capt. LSTM) and without (LSTM) access to a video frame. The numbers above each token show the models' losses when predicting particular tokens (heatmap normalized within an utterance). The mean loss M is also shown. The Captioning (Capt.) LSTM has better mean loss than the LSTM on most examples, and the word predictions for some visible objects are improved over the LSTM ("doggy," "ball" in the third row, etc.). The third to sixth examples are harder to interpret: the Capt. LSTM fails to make better word predictions for other visible objects ("ball" in fourth row and "car"). Finally, the last two examples mention objects that are not present in the image ("banana" and "bear"). Nevertheless, the word "banana" is more likely in the Capt. LSTM presumably due to the correlation with the visual context (kitchen background); on the contrary, in the last example, the prediction on the word "bear" that does not have a corresponding visual referent becomes worse.

help to shape the networks' behaviors, including sensitivity to category-conforming contexts and phenomena, such as determiner-noun agreement, although longer distance dependencies proved more difficult (e.g., subject-verb agreement). Second, the networks can also organize nouns into semantic categories, such as animals, body parts, and clothing, largely following a taxonomic organization mixed with some thematic influences. Finally, we found that introducing visual information brings an incremental improvement for predicting words in context, with relatively larger improvements for syntactic categories, such as nouns and verbs. However, the acquired linguistic representations in the LSTMs were similar regardless of whether they received visual information.

A distinguishing aspect of our work is using naturalistic, multimodal data from a single child. Elman's pioneering work (1989, 1990, 1991) showed how SRNs can learn meaningful syntactic and semantic representations without targeted inductive biases. The NLP community has continued this tradition, using modern successors of the SRN for modeling sequences (LSTMs, Transformers, etc.) trained on larger-scale written text corpora (Belinkov & Glass, 2019; Linzen & Baroni, 2021; Rogers, Kovaleva, & Rumshisky, 2021; Warstadt & Bowman, 2022). Moreover, neither synthetic nor written text is essential: networks can also learn useful syntactic and semantic representations when trained on the naturalistic, noisy data received by multiple children (Fourtassi, 2020; Huebner & Willits, 2018; Huebner et al., 2021). Our work takes a further step in demonstrating how the same types of regularities, although in more nascent forms, emerge from neural networks trained on the linguistic input received by just one child. Furthermore, we also provide an initial examination of what additionally can be learned when visual data are paired with the linguistic input, complementing previous work training vision-only models on SAYCam (Orhan et al., 2020; Zhuang et al., 2021).

By using data from just one child, we inevitably have less training data than previous studies with aggregate corpora. Unsurprisingly, data quantity impacts the acquisition of linguistic structure (Warstadt, Zhang, Li, Liu, & Bowman, 2020). The 225K tokens in our training set is a small fraction of a child's overall input. Assuming a child receives roughly 3M to 20M words per year (Appendix S1 of Dupoux, 2018), our training data are 0.5%–4% of the child's input in the first 2 years or even smaller fraction of this, considering that not all of the SAYCam tokens are words (e.g., punctuations). In contrast, BabyBERTa (Huebner et al., 2021) that was trained on 5M words (using AO-CHILDES; aggregated from multiple children and spanning a longer age range) achieved stronger performance on acceptability judgments (Fig. 5). More work is needed to understand the nature of these differences: these gaps may arise from differences in terms of data scale or data diversity due to more children across more ages and more environments. We see our method as a conservative approach, using real rather than proxy data available to one learner, that ensures models will not benefit from the additional diversity of aggregated data. Nonetheless, we see complementary value in both methodologies, trading off between data quantities and more realistic settings. We hope that the future will bring denser and longer-range datasets from individual children, mitigating these trade-offs and facilitating even more powerful studies of learnability.

Although we focused on the outcome of learning rather than the stages of learning—that is, we did not seek to build a model of cognitive development—it is still instructive to compare our findings to studies of language acquisition in children. We have demonstrated

that distributional information in the input to a child before 25 months of age is enough to support the formation of syntactic categories, including nouns and nonalternating transitive and intransitive verbs. Meanwhile, children's category structures develop at varying paces. For example, children at around 23 months can productively use novel nouns but not verbs, indicating a more well-formed grammatical category for nouns (compared to verbs) at this age (Olguin & Tomasello, 1993; Tomasello & Olguin, 1993). Our networks' failure to acquire more complex linguistic phenomena, in particular, subject-verb agreement, may also benefit from a parallel discussion with developmental work. English-speaking children have been reported to successfully produce subject-verb agreement markers between the ages of 2;2 and 3;10 (Brown, 1973). Given that the endpoint of our training data is 25 months, it may be the case that access to a child's linguistic input that extends beyond this timeframe is required. Furthermore, the comprehension of subject-verb agreement has been known to be delayed in English-speaking children (Johnson, de Villiers, & Seymour, 2005; Legendre et al., 2014). In this regard, our results provide a piece of supporting evidence speaking to the weakness of distributional cues for subject-verb agreement in early child-directed input.

Regarding semantic development, our results showed that the emergent semantic clusters of words correspond to real superordinate categories that children learn ("animal," "vehicle," etc.), although exactly when and how children learn these concepts is still a puzzle (Murphy, 2002). Infants can discriminate between visual exemplars of superordinate categories (animal vs. vehicle) in the first few months of life, with discrimination between more specific categories (Saint Bernard vs. Beagle) emerging later (Mandler & McDonough, 1993; Quinn, 2004). On the other hand, language seems to follow a different path: words for superordinate categories are acquired comparatively late relative to words for basic-level categories (Murphy, 2016). Additionally, the developmental timecourse of taxonomic relatedness, compared to more associative and thematic forms of relatedness, is still debated and seems to vary according to the task (Gelman & Markman, 1986; Markman & Hutchinson, 1984; Sloutsky, Yim, Yao, & Dennis, 2017; Unger, Savic, & Sloutsky, 2020; Unger & Fisher, 2021). Our results suggest that information regarding taxonomic (including superordinate) categories can be readily extracted from a small subset of the linguistic input to one child (up to age 3), as found in other modeling work using broader aggregate data (Sloutsky et al., 2017). It is thus unclear what underlies the differences between modalities and the late acquisition of some types of semantic and conceptual knowledge; future work using other multimodal models trained on SAYCam could potentially provide a unique lens into these questions.

Our work only scratches the surface of understanding what is learnable from a young child's experiences. SAYCam offers an unprecedented snapshot of three children's experiences, but it captures only a small fraction of their total linguistic input, preventing us from analyzing more complex linguistic phenomena (Belinkov & Glass, 2019; Linzen & Baroni, 2021; Rogers et al., 2021). Our multimodal training setup also does not capture the full richness of the multimodal signals that children may receive. Beyond imperfections in preprocessing (Section 2) and the inherent stochasticity in a child's gaze (Yu, Zhang, Slone, & Smith, 2021), the use of tokenized text rather than audio removes phonological or morphological cues, while also treating segmentation capabilities as given (Meylan & Bergelson, 2022). We mainly focused on linguistic analyses that are applicable to text-only

setups, because this enables us to better isolate the contribution of introducing multimodality. Nevertheless, a very important future direction is to investigate grounded semantics of the language, with multimodal neural networks like our captioning model or contrastive models, using relevant tasks, such as image-text matching or cross-modal forced-choice paradigms (Chrupała, Gelderloos, & Alishahi, 2017; Harwath et al., 2018; Khorrami & Räsänen, 2021; Kádár, Chrupała, & Alishahi, 2015; Lazaridou, Chrupała, Fernández, & Baroni, 2016; Nikolaus & Fourtassi, 2021; Vong & Lake, 2022). Moreover, we did not fully incorporate the temporal nature of a child's experience, both in how the videos were converted to still images (impeding learning of certain kinds of words that might require visuotemporal integration, e.g., "pick" and "take"; Ebert & Pavlick, 2020) and how networks were trained on the whole corpus simultaneously (one alternative, training networks on age-ordered data, can be found in Huebner & Willits, 2020). A future extension in the network architecture could incorporate the temporal structure of video frames, such as attention-based pooling or more generally video network architectures (Merkx, Frank, & Ernestus, 2019; Tran et al., 2018). Potentially, dialog models could also help in learning from interactive linguistic contexts. In addition to modeling the temporal structure, an even harder future challenge is limiting models to one pass through the data as a stricter criterion for learnability.

Finally, and perhaps most importantly, the networks learn passively from a child's fundamentally active and embodied experiences under the current setup. The networks cannot choose their own actions to take in the environment, do not have desires and goals, do not utilize social cues in support of learning, and do not realize that language can be a means of achieving what they want. In all of these ways, the types of neural networks considered here, even when scaled up, are far from understanding language in all the ways that people do (Lake and Murphy, 2021). Nevertheless, our results show that neural networks can acquire meaningful structures from a real snapshot of developmental experience. Stronger models, paired with denser and higher-resolution developmental snapshots, would undoubtedly lead to further discoveries.

Acknowledgments

We are grateful for Jeffrey Elman and his many contributions to cognitive science. Jeff published "Finding structure in time" (Elman, 1990) over 30 years ago, yet his article continues to guide cognitive science, natural language processing, and other fields today. We thank Gregory L. Murphy and three anonymous reviewers for feedback on earlier drafts of this article. We are also grateful for the volunteers who contributed to the SAYCam dataset (Sullivan et al., 2021) that made our article possible.

Notes

1 As in previous work, we draw parallels between emergent clusters of word embeddings and real-world categories ("animal," "vehicle," etc.). Importantly, however, these learned representations are quite limited in function and structure compared to

full-fledged human conceptual representations (Lake and Murphy, 2021). We elaborate on this point in the General Discussion.

2 Our code can be found on https://github.com/wkvong/multimodal-baby.

3 The SAYCam dataset can be accessed on https://nyu.databrary.org/volume/564. Access can be provided to academic investigators through the Databrary authorization process.

4 Although this interpolation procedure did not lead to time spans that were exactly aligned with each of the spoken utterances, the relative stability of visual information across seconds meant that the approximate alignment was still informative. We note that noise introduced at this step would lead to an underestimate, not an overestimate, of learnability.

5 The temporal order of utterances is not taken into account. They are also randomly ordered when presented to the network. So, the network treats each frame-utterance pair as an independent datapoint.

6 The hidden state and embedding sizes were not critical for our analyses; Smaller embedding dimensions led to degradation of performance on token prediction, but the qualitative conclusions of our analyses remained unchanged.

7 In natural language processing, perplexity is a measure of how well a predicted distribution matches the ground-truth one-hot token distribution, defined as $\frac{1}{\tilde{p}(\mathrm{y})}$, where $\tilde{p}(\mathrm{y})$ is the predicted probability of the ground-truth token y. For a corpus consisting of n tokens, the perplexity is defined as $\exp(\frac{1}{n}\sum_{i=1}^{n} -\log \tilde{p}(\mathrm{y}_i))$, where y_i is the i-th token. The lower the perplexity, the better.

8 In order to make perplexity as comparable as possible across LSTM and CBOW, all these numbers exclude Start-Of-Sequence (SOS) and End-Of-Sequence (EOS) tokens appended to the starts and ends of utterances, so they are evaluated on the same set of tokens.

9 Note that it has been shown that small contexts primarily encode syntactic aspects over thematic ones (Chang & Deák, 2020; Huebner and Willits, 2018).

10 See Appendix A.3 for details of how we classify the transitivity of verbs.

11 While we used word embeddings to conduct these analyses, mean hidden vectors across the dataset (the approach used by Elman, 1990) yield similar results.

12 See Appendix A.3 for details of how we select nouns and label their semantic categories.

13 CBOW results are shown in Figure 9 in the Appendix; there are also many identifiable clusters like body parts and clothing, but many others are less clear than clusters from the LSTM.

14 This approach is similar to the category distinction test for masked language models in Kim and Smolensky (2021).

15 The Captioning LSTM always needs an image input, so we used the mean image frame of the training set in this evaluation. Of course, this mean image does not specifically relate to the candidate sentences in the evaluation. As shown in Fig. 5, its performance is not substantially different from the language-only LSTM.

16 n-gram models are simple language models based on token counts in a corpus. An n-gram is n consecutive tokens. The unigram model is based on counts of individual

token, without considering any context. The bigram model is based on counts of token pairs occurring together, and so on. We tried larger n-gram models for the acceptability analysis, but they performed similarly to the bigram model due to data sparsity and their back-off mechanism. The back-off mechanism of an n-gram model is that when the n-gram has 0 count in the training set, in order to avoid 0 probability, the model will try using the probability of the shorter $(n-1)$-gram, and so on.

17 In fact, the AO-CHILDES Transformer trained on more data also shows comparatively worse performance on this test compared to other tests.

18 https://github.com/eminorhan/baby-vision

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

Belinkov, Y., & Glass, J. R. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72.

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives*, *14*(3), 142–149.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*, 3253–3258.

Bergelson, E., & Swingley, D. (2015). Early word comprehension in infants: Replication and extension. *Language Learning and Development*, *11*, 369–380.

Berzak, Y., Barbu, A., Harari, D., Katz, B., & Ullman, S. (2015). Do you see what I mean? Visual resolution of linguistic ambiguities. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1477–1487). Lisbon, Portugal: Association for Computational Linguistics.

Bird, S., Loper, E., & Klein, E. (2009). Natural language processing with Python. O'Reilly Media Inc.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Brown, R. S. (1973). *A first language: The early stages*. Harvard University Press.

Chang, L., & Deák, G. O. (2020). Adjacent and non-adjacent word contexts both predict age of acquisition of English Words: A distributional corpus analysis of child-directed speech. *Cognitive Science*, *44*(11), e12899.

Christie, G., Laddha, A., Agrawal, A., Antol, S., Goyal, Y., Kochersberger, K., & Batra, D. (2016). Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1493–1503). Austin, TX: Association for Computational Linguistics.

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 613–622). Vancouver, Canada: Association for Computational Linguistics.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Ebert, D., & Pavlick, E. (2020). A visuospatial dataset for naturalistic verb learning. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics* (pp. 143–153). Barcelona, Spain (Online): Association for Computational Linguistics.

Elman, J. L. (1989). Representation and structure in connectionist models. *Technical report*.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis* (pp. 1–32). Wiley-Blackwell.

Fourtassi, A. (2020). Word co-occurrence in child-directed speech predicts children's free word associations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 49–53). Online: Association for Computational Linguistics.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data*. *Journal of Child Language*, *44*, 677–694.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank Project*. MIT Press.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 146–162.

Harwath, D. F., Recasens, A., Surís, D., Chuang, G., Torralba, A., & Glass, J. R. (2018). Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*, *128*, 620–641.

Hill, F., Tieleman, O., von Glehn, T., Wong, N., Merzic, H., & Clark, S. (2021). Grounded language learning fast and slow. In *International Conference on Learning Representations*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 624–646). Association for Computational Linguistics.

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, *9*, 133.

Huebner, P. A., & Willits, J. A. (2020). Order matters: Developmentally plausible acquisition of lexical categories. In *CogSci*.

Huebner, P. A., & Willits, J. A. (2021). Using lexical context to discover the noun category: Younger children have it easier. In K. D. Federmeier & L. Sahakyan (Eds.), *The context of cognition: Emerging perspectives*, volume 75 of *Psychology of learning and motivation* (pp. 279–331). Academic Press.

Iki, T., & Aizawa, A. (2021). Effect of visual extensions on natural language understanding in vision-and-language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 2189–2196). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Johnson, V. E., de Villiers, J. G., & Seymour, H. N. (2005). Agreement without understanding? The case of third person singular/s. *First Language*, *25*(3), 317–330.

Kádár, Á., Chrupała, G., & Alishahi, A. (2015). Lingusitic analysis of multi-modal recurrent neural networks. In *Proceedings of the Fourth Workshop on Vision and Language* (pp. 8–9). Lisbon, Portugal: Association for Computational Linguistics.

Khorrami, K., & Räsänen, O. J. (2021). Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? – A computational investigation. *ArXiv*, *abs/2109.14200*.

Kim, N., & Smolensky, P. (2021). Testing for grammatical category abstraction in neural language models. In *Proceedings of the Society for Computation in Linguistics 2021* (pp. 467–470). Association for Computational Linguistics.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*, *130*(2), 401–431.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lazaridou, A., Chrupała, G., Fernández, R., & Baroni, M. (2016). Multimodal semantic learning from child-directed input. In *North American Chapter of the Association for Computational Linguistics*.

Legendre, G., Culbertson, J., Culbertson, J., Zaroukian, E. G., Hsin, L. B., Barrière, I., & Nazzi, T. (2014). Is children's comprehension of subject-verb agreement universally late? Comparative evidence from French, English, and Spanish. *Lingua*, *144*, 21–39.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212.

Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates Publishers.

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, *8*, 291–318.

Markman, E. M. (1989). *Categorization and naming in children*. Cambridge, MA: MIT Press.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, *16*, 1–27.

Merkx, D., Frank, S., & Ernestus, M. (2019). Language learning using speech to image retrieval. In *Interspeech* (pp. 1841–1845).

Meylan, S. C., & Bergelson, E. (2022). Learning through processing: Toward an integrated approach to early word learning. *Annual Review of Linguistics*, *8*, 77–99.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR*.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L. (2016). Explaining the basic-level concept advantage in infants… or is it the superordinate-level advantage? *Psychology of Learning and Motivation*, *64*, 57–92.

Nikolaus, M., & Fourtassi, A. (2021). Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 200–210). Association for Computational Linguistics.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, *8*, 245–272.

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., & Lin, H. (Eds.), *Advances in neural information processing systems*, volume 33 (pp. 9960–9971). Curran Associates, Inc.

Pérez-Mayos, L., Ballesteros, M., & Wanner, L. (2021). How much pretraining data do language models need to learn syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1571–1582). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*, 306–338.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Quinn, P. C. (2004). Development of subordinate-level categorization in 3- to 7-month-old infants. *Child Development*, *75*, 886–899.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

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

Shi, F., Mao, J., Gimpel, K., & Livescu, K. (2019). Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1842–1861). Florence, Italy: Association for Computational Linguistics.

Siddharth, N., Barbu, A., & Siskind, J. M. (2014). Seeing what you're told: Sentence-guided activity recognition in video. *2014 IEEE Conference on Computer Vision and Pattern Recognition*.

Sloutsky, V. M., Yim, H., Yao, X., & Dennis, S. J. (2017). An associative account of the development of word learning. *Cognitive Psychology*, *97*, 1–30.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E. H., & Frank, M. C. (2021). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, *5*, 20–29.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., & Metzler, D. (2021). Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations (ICLR)*.

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism & Mass Communication Quarterly*, *30*, 415–433.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*, 172–175.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*(4), 432–444.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, *8*, 451–464.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Unger, L., & Fisher, A. V. (2021). The emergence of richly organized semantic knowledge from simple statistics: A synthetic review. *Developmental Review*, *60*, 100949.

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

van der Maaten, L., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in neural information processing systems*, volume 30, 6000–6010. Curran Associates, Inc.

Vong, W. K., & Lake, B. M. (2022). Cross-situational word learning with multimodal neural networks. *Cognitive Science*, *46*(4), e13122.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. *ArXiv*, *abs/2208.07998*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

Warstadt, A., Zhang, Y., Li, X., Liu, H., & Bowman, S. R. (2020). Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 217–235). Association for Computational Linguistics.

Waterfall, H. R., Sandbank, B., Onnis, L., & Edelman, S. (2010). An empirical generative framework for computational modeling of language acquisition. *Journal of Child Language*, *37*(3), 671–703.

Wojcik, E. H. (2018). The development of lexical-semantic networks in infants and toddlers. *Child Development Perspectives*, *12*, 34–38.

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yu, C., Zhang, Y., Slone, L. K., & Smith, L. B. (2021). The infant's view redefines the problem of referential uncertainty in early word learning. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(52), e2107019118.

Yu, H., Siddharth, N., Barbu, A., & Siskind, J. M. (2015). A compositional framework for grounding language inference, generation, and acquisition in video. *Journal of Artificial Intelligence Research*, *52*, 601–713.

Yun, T., Sun, C., & Pavlick, E. (2021). Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 4357–4366). Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(3), e2014196118.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix