Finding Structure in One Child's Linguistic Experience

Wentao Wang, **Wai Keen Vong**, **Najoung Kim**, and **Brenden M. Lake**
Center for Data Science, New York University

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 Elman's (1990) 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, and which kinds may benefit from stronger inductive biases or richer sources of data.

1 Introduction

In the first three years of life, children's linguistic development progresses rapidly. Young children begin understanding words at around 6 months (Tincoff and Jusczyk, 1999, 2012; Bergelson and Swingley, 2012, 2015). The vocabulary that they can comprehend and produce increases gradually until around 12–14 months, at which a non-linear 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 and 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 without language-specific inductive biases on this data and evaluate what these models learn (e.g., Orhan et al. 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 et al., 2015; Sullivan et al., 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 and Willits, 2018; Warstadt and Bowman, 2022) without the aid of language-specific inductive biases, social cognition abilities, and aspects of world knowledge that are thought to play central roles (Markman, 1989; Bloom, 2000; 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 to 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 and Schmidhuber, 1997) and Continuous Bag-Of-Words (CBOW; Mikolov et al., 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 (Elman, 1990; Perfors et al., 2011; Abend et al., 2017; Huebner and Willits, 2018; Huebner et al., 2021, *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 LSTM (Hochreiter and Schmidhuber, 1997) and Transformer (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 and 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 (Waterfall et al., 2010; Perfors et al., 2011; Abend et al., 2017). 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, labels, or language-specific inductive biases.

From our simulations and analyses, we have both positive and negative findings regarding learnability. 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 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 is a subset of SAYCam (Sullivan et al., 2021), a dataset consisting of egocentric head-mounted camera recordings of 3 very young, English-speaking children.[2] Each child's recordings are recorded at regular intervals (several hours each week) for around 2 years starting from 6–8 months of age. However, out of the 3 children, only one (labeled as baby S) had a large proportion of his naturalistic speech input transcribed (spanning 6–25 months of

[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] 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.

	Train	Validation	Test
Number of utterances	33,737	1,874	1,875
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%

Table 1: **Statistics of SAYCam-S.**

age), making baby S the choice for 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 and 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.[3] We filter the 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 6.4s of its time span.

The dataset is randomly split into training, validation and test sets (90%/5%/5% of all utterances, respectively).[4] 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 2,350. 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 Figure 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

3.1 Language-only networks

We use two kinds of networks to encode the language input: single-layer uni-directional LSTM (Hochreiter and Schmidhuber, 1997), which is a variant of Recurrent Neural Network (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. 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, and so on). Instead, we utilize 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.

Figure 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.

[3]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.

[4]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.

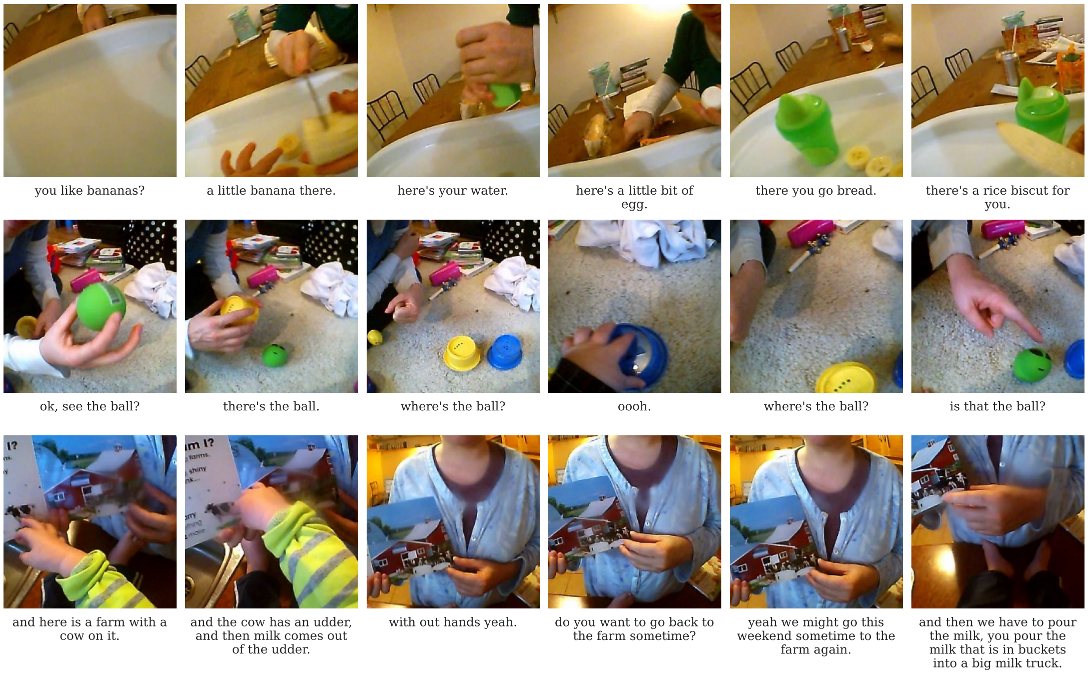


Figure 1: **Example frames and their corresponding utterances.** Each row is a different scene: eating breakfast, playing a game with a ball, and reading a farm-themed picture book. Unlike common image-text datasets in machine learning, the utterances only loosely align to the frames. For instance, the 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.

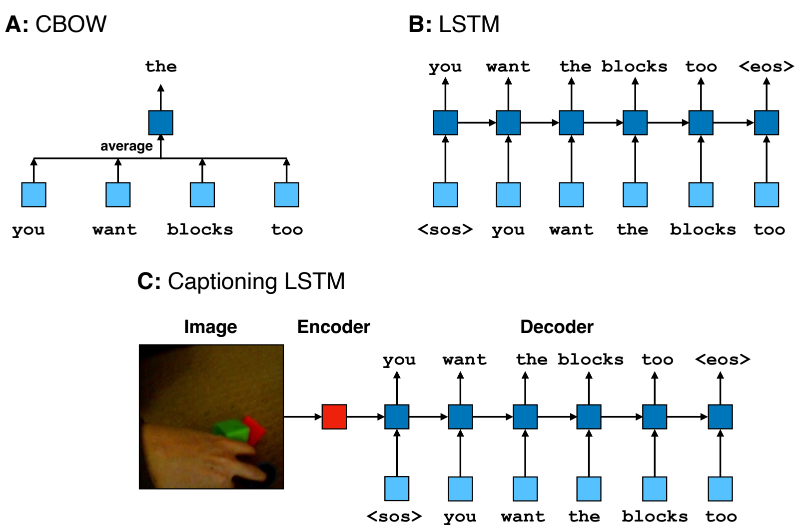


Figure 2: **The three neural network architectures.** (a) The CBOW network predicts a missing word given a surrounding context of fixed size. The LSTM (b) and Captioning LSTM (c) networks both predict the next word given a sequence of previous words (additionally a corresponding image for the Captioning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

The dimensions of the hidden states and the word embeddings are both 512.[5] When predicting the next token, the LSTM assigns a probability distribution over all tokens in the vocabulary.

[5]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.

Figure 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 are 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. 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.[6] 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 3 runs with different random seeds.[7] 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 to 4 tokens on both sides of the predicted token and found that a context window containing only 1 token on both sides performed best.[8]

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 and Lake, 2022). Multimodal learning has been shown to help resolve ambiguities when only linguistic information is present (Berzak et al., 2015; Christie et al., 2016), induce constituent structures (Shi et al., 2019), and ground events described in language to video (Siddharth et al., 2014; Yu et al., 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 Figure 2(c). We use a Convolutional Neural Network (CNN) as our vision encoder (specifically, ResNeXt-50 32x4d; Xie et al., 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.

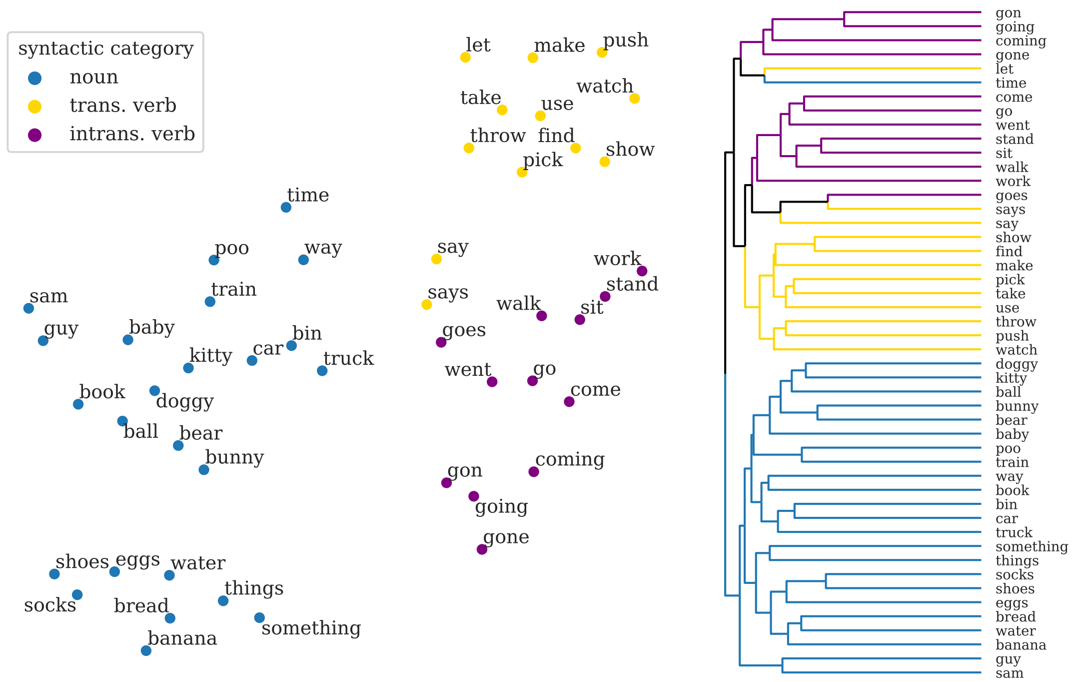
The perplexity of our Captioning LSTM was 22.10 ($SD = 0.20$) averaged over 3 runs, which was incrementally lower than the language-only LSTM, suggesting a minor benefit of information from the additional visual modality. Noise in the alignment between the visual and language streams likely damped the size of the improvement. We discuss this issue further in the context of the limitations of the multimodal objective in the General Discussion.

---

[6] 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.

[7] 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.

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



(a) t-SNE (b) Dendrogram clustering

Figure 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.

4 Results

4.1 Learning from language only

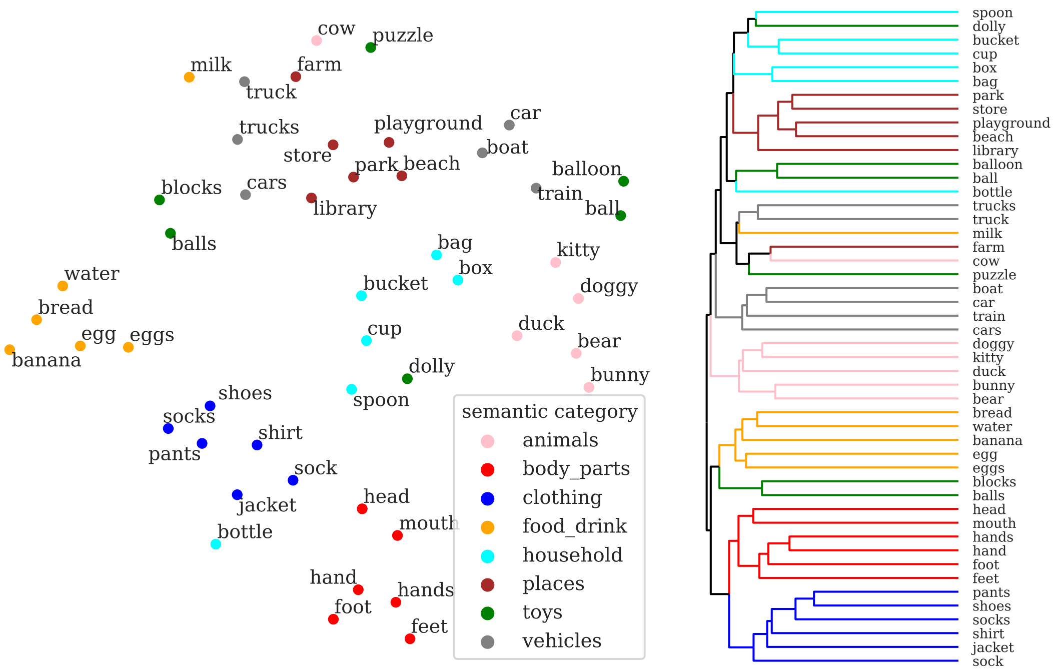
4.1.1 Syntactic and semantic categories

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, reporting the results for the LSTM in the main text and the corresponding results for CBOW 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 Figure 3; we visualize the LSTM's word embeddings using t-SNE (van der Maaten and Hinton, 2008) and a dendrogram for the most frequent 24 nouns—12 transitive verbs, and 12 intransitive verbs that are unambiguous in their transitivity[9] (see Figure 8 in the Appendix for CBOW results). Both the t-SNE and dendrogram use cosine-based metrics between word embeddings.[10] Furthermore, Figures 12 and 13 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.

[9] See Appendix A.3 for details of how we classify the transitivity of verbs.

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



(a) t-SNE

(b) Dendrogram clustering

Figure 4: **Clustering LSTM's word embeddings for semantic categories.** Again, both plots use cosine measures in Figure 3. We present the most frequent 6 words from 8 different categories. Most distinct clusters clearly correspond to semantic categories.

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 et al., 2016).[11] We exclude categories having less than 6 unambiguous words from our analysis. As can be seen from Figure 4, there are several visually identifiable clusters that correspond to different semantic categories.[12] Note that while Elman (1990) found a clear animate versus inanimate distinction among nouns, we did not find such a salient distinction (see Figure 15 in 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 scene in the training data; these words co-occur in a scene where the parent is reading a farm-themed picture book, illustrated in the third row of Figure 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.[13] 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 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 3 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

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

[12] 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.

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

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

Table 2: **Examples of clozes and the networks' predictions.** 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.

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., 2020a). The minimal pair approach asks models to judge which of 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 7 different linguistic phenomena; 8 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[14]), two baseline N-gram language models based on statistics of the training set (unigram and bigram language models[15]),

[14] 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 Figure 5, its performance is not substantially different from the language-only LSTM.

[15] 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.)

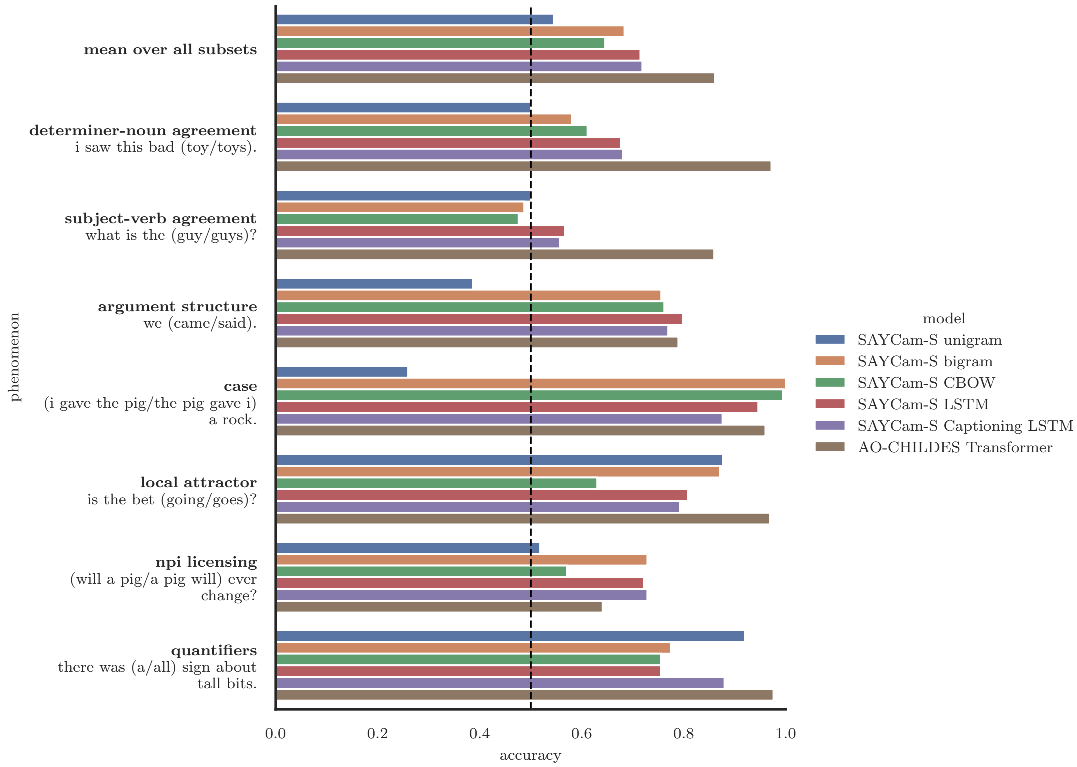


Figure 5: **Mean linguistic acceptability test accuracy over subsets for each network and linguistic phenomenon.** The top group of bars is the mean over all subsets, and each of the remaining groups is the mean over the subsets corresponding to specific linguistic phenomena. Each label for a phenomenon 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.

and a strong Transformer model (pre-trained weights from Huebner et al. 2021 trained on AO-CHILDES which aggregates data from many children). The results are summarized in Figure 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 and Leonard, 2018) and Transformers (Tay et al., 2021; Pérez-Mayos et al., 2021) with modest amounts of training data in general have increased difficulty with longer-distance dependencies.[16] 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.

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

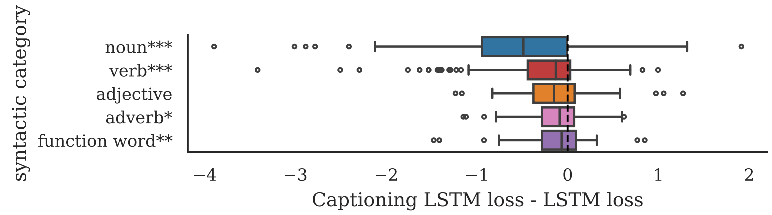


Figure 6: **Type-level loss difference between language-only LSTMs and Captioning LSTMs on the validation set.** Losses are means over all occurrences of the word type and all 3 runs for each architecture. *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$. More negative values on the x-axis indicate more improvement with added visual information. See Table 7 in Appendix for detailed t-test results.

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 Figure 6. The improvements for most syntactic categories are statistically significant (Table 7 in 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 et al., 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 Figure 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 et al., 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 and Aizawa, 2021; Yun et al., 2021).

Figure 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 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 due to the correlation with the visual context; on the contrary, in the last example, the prediction on the word "bear" that does not have a corresponding visual referent becomes worse.

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, we find that neural networks learning exclusively from developmentally plausible data can differentiate words in different syntactic categories. These categories help to shape the networks' behaviors, in predicting a token's category based on context and in acquiring sensitivity to 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 it 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 Simple Recurrent Networks (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 and Glass, 2019; Rogers et al., 2021; Linzen and Baroni, 2021; Warstadt and 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 (Huebner and Willits, 2018; Huebner et al., 2021; Fourtassi, 2020). 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 is 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 et al., 2020b). The 225K tokens in our training set is a small fraction of a child's overall input (roughly 0.5% to 4% of the child's input in the first 2 years), assuming a child receives roughly 3M to 20M words per year (Dupoux, 2018, Appendix S1).[17] In contrast, BabyBERTa (Huebner et al., 2021) was trained on 5M words (using AO-CHILDES; aggregated from multiple children and spanning a longer age range) and achieved stronger performance on acceptability judgments (Figure 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 non-alternating 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 (Tomasello and Olguin, 1993; Olguin and Tomasello, 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 et al., 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.

[17] Additionally, not all of the SAYCam tokens are words (e.g., punctuation) and thus the fraction is reduced further.

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 and 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 (Markman and Hutchinson, 1984; Gelman and Markman, 1986; Sloutsky et al., 2017; Unger et al., 2020; Unger and 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 knowledge; 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 training larger and more sophisticated networks (e.g., Transformers; Vaswani et al., 2017) or analyzing more complex linguistic phenomena (Belinkov and Glass, 2019; Rogers et al., 2021; Linzen and Baroni, 2021). The challenges of training multimodal models are particularly noteworthy. Beyond imperfections in pre-processing (Section 2) and inherent stochasticity in a child's gaze (Yu et al., 2021), using tokenized text rather than audio removes phonological or morphological cues, while also treating segmentation capabilities as given (Meylan and Bergelson, 2022). We mainly focused on linguistic analyses that are applicable to text-only setups, because this enables us to study the contribution of introducing multimodality. 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 (Kádár et al., 2015; Lazaridou et al., 2016; Chrupała et al., 2017; Harwath et al., 2018; Khorrami and Räsänen, 2021; Nikolaus and Fourtassi, 2021; Vong and 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 and 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 and 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 et al., 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 must learn passively from a child's fundamentally active and embodied experiences. 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.

References

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

Belinkov, Y. and 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. and 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. and 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., and Ullman, S. (2015). Do You See What I Mean? Visual Resolution of Linguistic Ambiguities. In *EMNLP*.

Bird, S., Loper, E., and Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

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

Brown, R. S. (1973). *A First Language: The Early Stages*. Harvard University Press.

Chang, L. and 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., and 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*, pages 1493–1503, Austin, Texas. Association for Computational Linguistics.

Chrupała, G., Gelderloos, L., and 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)*, pages 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. and Pavlick, E. (2020). A Visuospatial Dataset for Naturalistic Verb Learning. In *STARSEM*.

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*, pages 1–32. Wiley-Blackwell.

Fourtassi, A. (2020). Word Co-occurrence in Child-directed Speech Predicts Children's Free Word Associations. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

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

Gelman, S. A. and 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., and 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., and Clark, S. (2021). Grounded language learning fast and slow. In *International Conference on Learning Representations*.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9:1735–1780.

Honnibal, M. and 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., and Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Huebner, P. A. and Willits, J. A. (2018). Structured Semantic Knowledge Can Emerge Automatically from Predicting Word Sequences in Child-Directed Speech. *Frontiers in Psychology*, 9.

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

Huebner, P. A. and Willits, J. A. (2021). Using lexical context to discover the noun category: Younger children have it easier. In Federmeier, K. D. and Sahakyan, L., editors, *The Context of Cognition: Emerging Perspectives*, volume 75 of *Psychology of Learning and Motivation*, pages 279–331. Academic Press.

Iki, T. and Aizawa, A. (2021). Effect of Visual Extensions on Natural Language Understanding in Vision-and-Language Models. In *EMNLP*.

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

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

Khorrami, K. and 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. and Smolensky, P. (2021). Testing for grammatical category abstraction in neural language models. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 467–470, Online. Association for Computational Linguistics.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, 2.

Lake, B. M. and Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*.

Landauer, T. K. and 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., and 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., and 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. and Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Linzen, T. and 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*, page 692 – 697.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates Publishers, 3rd edition.

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

Markman, E. M. (1989). *Categorization and Naming in Children*. MIT Press, Cambridge, MA.

Markman, E. M. and 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., and Ernestus, M. (2019). Language learning using speech to image retrieval. In *Interspeech*, pages 1841–1845.

Meylan, S. C. and 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., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press, Cambridge, MA.

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. and 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*, pages 200–210, Online. Association for Computational Linguistics.

Olguin, R. and 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., and Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9960–9971. Curran Associates, Inc.

Pérez-Mayos, L., Ballesteros, M., and Wanner, L. (2021). How much pretraining data do language models need to learn syntax? In *EMNLP*.

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

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and 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., and 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., and 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., and Livescu, K. (2019). Visually Grounded Neural Syntax Acquisition. In *ACL*.

Siddharth, N., Barbu, A., and Siskind, J. M. (2014). Seeing What You're Told: Sentence-Guided Activity Recognition in Video. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 732–739.

Sloutsky, V. M., Yim, H., Yao, X., and 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., and 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., and 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. and Jusczyk, P. W. (1999). Some Beginnings of Word Comprehension in 6-Month-Olds. *Psychological Science*, 10:172 – 175.

Tincoff, R. and Jusczyk, P. W. (2012). Six-Month-Olds Comprehend Words That Refer to Parts of the Body. *Infancy : the official journal of the International Society on Infant Studies*, 17 4:432–444.

Tomasello, M. and 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., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Unger, L. and Fisher, A. V. (2021). The Emergence of Richly Organized Semantic Knowledge from Simple Statistics: A Synthetic Review. *Developmental review : DR*, 60.

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

van der Maaten, L. and 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., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vong, W. K. and Lake, B. M. (2022). Cross-Situational Word Learning With Multimodal Neural Networks. *Cognitive science*, 46 4:e13122.

Warstadt, A. and 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., and Bowman, S. R. (2020a). 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., and Bowman, S. R. (2020b). 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)*, pages 217–235, Online. Association for Computational Linguistics.

Waterfall, H. R., Sandbank, B., Onnis, L., and 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., and He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*.

Yu, C., Zhang, Y., Slone, L. K., and 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.

Yu, H., Siddharth, N., Barbu, A., and Siskind, J. M. (2015). A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. *J. Artif. Intell. Res.*, 52:601–713.

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

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and 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.

A Appendix

A.1 Dataset Details

SAYCam (Sullivan et al., 2021) is a longitudinal dataset consisting of egocentric head-mounted camera recordings from 3 children (S, A and Y), whose recordings span the ages 6–30, 8–31, and 7–24 months respectively. Recordings took place for a few hours each week over this course, amounting to 100–200 hours of recorded video data per child and more than 415 hours in total. As mentioned in Section 2, we only use the data from baby S since their videos had the largest proportion of speech transcribed. The speech transcribed for this child spans 6–25 month of age. Each transcript contains the relevant information for our purposes, including the utterances, the speaker and the time of the utterance (in seconds).

Preprocessing of transcripts. There was considerable noise in the original transcripts, requiring a number of preprocessing steps before feeding them as input to our networks. Some of these issues included very long annotations of multiple sentences, sometimes spanning minutes of video, and inconsistencies across transcripts. To resolve the first issue regarding long annotations, we use spaCy (Honnibal and Montani, 2017) to split annotated utterances into shorter sentences, which are the utterances we actually use. As we mentioned in Section 2, we label the time span of each utterance by linearly interpolating (i.e., evenly segmenting) the time span of the original transcript. We filtered these utterances, retaining only those from either parent, which comprised the majority of the child-directed speech. As we also mentioned in Section 2, we excluded child-produced utterances to focus solely on the data a child receives, meeting our goal of investigating what can be learned from the input to a child. Additionally, many of the transcribed child utterances, especially earlier in language development, are not very informative. Utterances from people other than the parents are rare. The second issue is also mitigated by the spaCy tokenizer (Honnibal and Montani, 2017). For example, it separates the "i'm" into "i" and "'m", and "im" into "i" and "m", so that the model can recognize the same "i" across inconsistent transcripts. Of course, this is still imperfect, and we leave further improvements for future work. All transcripts were lowercased. When presented to the network, out-of-vocabulary tokens in utterances are replaced by `<UNK>`, and utterance lengths are truncated to at most 25 tokens.

Preprocessing of video frames. The original resolution of video frames from SAYCam are 640×480. In order to more closely mimic the view from the child and fit the input shape of our pretrained ResNeXt network (Orhan et al., 2020), we first resized the minor edge to 256 and then applied a 224×224 square crop centered at 16 pixels lower the center of each original frame. For each utterance, we extracted multiple video frames using this procedure at a rate of 5fps starting from the beginning of its time span until reaching the end of the time span or 32 extracted frames (6.4s of video). (We wanted to pick a reasonable number of temporal frames that were a power of 2 and where the visual content was mostly similar within the time span. 5fps was based on how Orhan et al. (2020) sampled the frames for their self-supervised training.)

A.2 Network and Training Configuration

Network Configuration. For all networks, we use embedding and hidden size 512. For the LSTM and the Captioning LSTM, we tie weights in the word embeddings with weights in the output layer and add bias terms with their output layers. For the CBOW, we do not tie the weights in the input and output embedding matrices, nor do we add bias terms.

For the LSTM, the starting hidden and cell states at the beginning of the sequence are initialized to all zeros. For the Captioning LSTM, we add a linear adapter layer on top of the vision encoder to project the visual representation and this projected representation is used to initialize the hidden and cell states of the uni-directional LSTM. We freeze the stem of the vision encoder and only train the adapter and LSTM. When training the network, we randomly sample a frame from the multiple frames aligned with the utterance, applied data augmentation, and yield an example pair $(\text{frame}, \text{utterance})$. The data augmentations we applied are the following (in PyTorch):

```
transforms.Compose([
    transforms.RandomResizedCrop((224, 224), scale=(0.2, 1.)),
    transforms.RandomApply([GaussianBlur([.1, 2.])], p=0.5),
    transforms.RandomHorizontalFlip(),
    transforms.ToTensor(),
    transforms.Normalize([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]),
```

```
])
```

Note this is the same set of augmentations as in the codebase of Orhan et al. (2020)[18], with the exception of ColorJitter as it breaks the correspondence between color words and color in images.

Training Configuration. Weights of all networks are randomly initialized by the default setting of PyTorch. Specifically, the weights of the LSTM and the output layer of the CBOW are initialized from Uniform($-\sqrt{1/d}, \sqrt{1/d}$) where d is the dimension 512, and the weights of the embeddings are initialized from Normal$(0, 1)$. For training the LSTM and the Captioning LSTM, we use batch size 16, initial learning rate 6×10^{-3}, and dropout on the input word embeddings with dropout rate 0.5. For training the CBOW, we use batch size 8, initial learning rate 3×10^{-3}, and dropout on the output embeddings with dropout rate 0.1. For all networks, we use the AdamW optimizer and apply weight decay of 0.04. For the LSTMs, we apply learning rate scheduling by reducing the learning rate by a factor of 10 when the validation loss has not improved across consecutive 5 epochs (same for CBOW, but with a 2 epoch threhold). The loss for a batch of utterances is the mean cross-entropy across all tokens. We apply early-stopping by training the network until convergence and selecting the checkpoint with the lowest loss on the validation set. All the hyperparameters are also tuned toward this validation loss. We trained each network with 3 different random seeds.

The performance of our networks measured in perplexities and the number of trained epochs for each selected best checkpoints are shown in Table 3.

Model	perplexity (SD)	number of trained epochs
CBOW	22.20 (0.01)	{31, 65, 58}
LSTM	24.80 (0.21)	{29, 38, 28}
Captioning LSTM	22.10 (0.20)	{29, 42, 38}

Table 3: Token prediction perplexities of networks on the validation set, and the numbers of trained epochs for selected best checkpoints of 3 runs. In order to make comparison across uni-directional networks and CBOW, we report perplexities excluding both the SOS and the EOS token. Perplexity numbers are the means of 3 runs, and numbers in the bracket are the standard deviations.

A.3 Selection and Categorization of Visualized Words

In this section, we describe the process of selecting and categorizing the words (nouns and verbs) that are visualized in our figures for syntactic (e.g., Figure 3) and semantic (e.g., Figure 4) categories. We first prepared semantic categories of nouns and syntactic categories of verbs: For nouns, we considered semantic categories from WordBank (Frank et al., 2016): sounds, animals, vehicles, toys, food&drink, clothing, body parts, household, furniture&rooms, outside, places, people, games&routines; for verbs, we considered these syntactic categories: transitive, ditransitive, intransitive, transitive/intransitive (which means the verb can be either transitive or intransitive), special (special verbs including be-verbs and modal verbs). We went through the most frequent words in our vocabulary by sorting them in descending order of their frequencies in the training set, and stopped at words with frequency 24 (due to limited time and labor). For each word we encountered, we did our best to classify it into the proper category, using examples from the dataset when needed. We excluded words having any of the following properties: 1) not a common word in the category (e.g., "marmite", "sam"), 2) ambiguous in its category (e.g., "chicken", "breaky", "painting"), and 3) referring to the categories themselves (e.g., "animals", "toys", "food"). For semantic categories, we decided to exclude 5 semantic subcategories (sounds, furniture&rooms, outside, people, games&routines) because we found they do not form coherent category structures or they did not contain enough words. For syntactic categories of verbs, we decided to only include transitive and intransitive verbs. In Table 4, we list samples of words that we excluded following the process described above.

A.4 Additional Clustering Figures

In this section, we include plots demonstrating that the learned networks are sensitive to other kinds of syntactic and semantic structure. First, we show additional t-SNE and dendrogram plots for the CBOW network showing that it can also differentiate nouns vs. verbs (Figure 8),

[18] `https://github.com/eminorhan/baby-vision`

POS	Category	Excluded Words
noun	sounds*	boop bloop ruff ya blo mmkay bop nom quack vroom boom
	animals	marmite chicken animals
	toys	toys toy book books bubbles dummy marker pen
	food & drink	food breakfast breaky chicken
	clothing	clothes nappy backpack blanket
	furniture & rooms*	computer
	outside*	sand flowers flower tree trees sun rocks
	places	house room
	people*	people sam guy toby
	games & routines*	game nap breaky
verb	transitive verb	painting
	ditransitive verb*	put give putting
	(in)transitive verb*	want see get know look like think try play read got end start
	special verb*	's is do are can have 're s done be did 'm 'll will wanna need

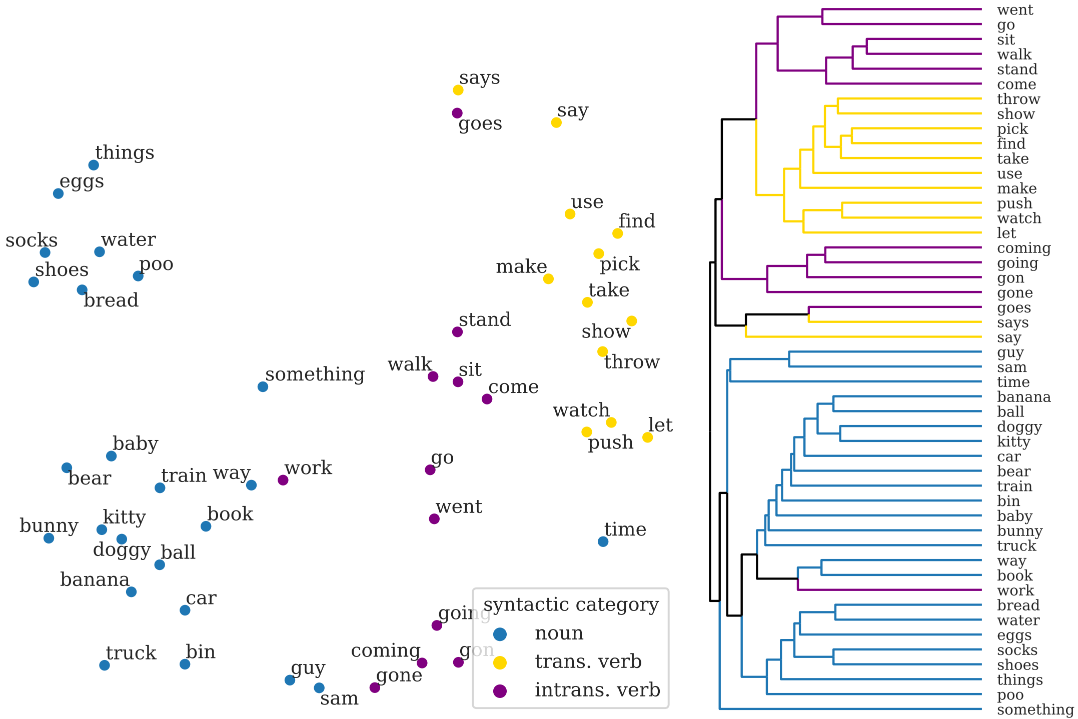
Table 4: Sample words we excluded, arranged into categories we considered they are closest to (for included categories) or they are in (for excluded categories, marked with *). To illustrate our process to determine whether to include or exclude a word, here is an example: for the word "marmite", we searched on the internet and found it is "a British savoury food spread", but examples in the dataset showed that it is the name of one of the family's cats; however, we still excluded this word because it is not a common word for "animals".

and also different semantic categories (Figure 9). Second, we also show additional t-SNE and dendrogram plots for the Captioning LSTM showing that its representations for words do not change much when given the images (Figure 10 and 11). Then, Figures 12, 13 and 14 show t-SNE and dendrogram plots that include words from additional syntactic categories (adjectives and adverbs) for the LSTM, CBOW and Captioning LSTM networks respectively, showing that all networks form clusters that correspond to each kind of syntactic category. Finally, Figure 15 presents a t-SNE plot showing different word embeddings from the LSTM network colored by animacy.

A.5 Cloze Test Details

As described in the main paper, we create clozes from utterances in the validation set. We filtered out clozes that contained less than two words, or occurred in the training set. We identify each token that is a noun or verb by using the POS tags labeled by Stanza POS tagger (Qi et al., 2020), and build the vocabulary of word fillers (nouns and verbs to fill into the clozes) by using every word that has its most frequent syntactic category as noun or verb, and is not ambiguous in its syntactic category (≥90% of its occurrences are with its most frequent syntactic category). This resulted in an evaluation set containing 2412 clozes, with 848 (35%) for nouns and 1564 (65%) for verbs, and vocabulary of word fillers containing 1439 words (1040 nouns and 399 verbs). In addition to the language-only models, we also evaluated the Captioning LSTMs by providing them the paired image frames and found they achieve 97.83% (SD = 0.10% over 3 runs), which is close to the result of language-only LSTMs. However, we noticed that many clozes have original words that are atypical nouns and verbs, such as be-verbs, modal verbs, quantifiers, words ambiguous in their part-of-speech, or `<UNK>` token. As a robustness check, we re-ran our cloze analysis after filtering out these clozes and excluding these atypical words. This left 1682 clozes, with 795 (47%) for nouns and 887 (53%) for verbs, and 1406 words in the filling word vocabulary (1034 nouns and 372 verbs). On this filtered set, our language-only LSTMs achieve 97.44% (SD = 0.10%) accuracy, CBOWs achieve 90.35% (SD = 0.28%) accuracy, and Captioning LSTMs achieve 97.42% (SD = 0.34%) accuracy. These accuracies are still high and similar to the results from the unfiltered set, suggesting that our results are robust to differences in vocabulary.

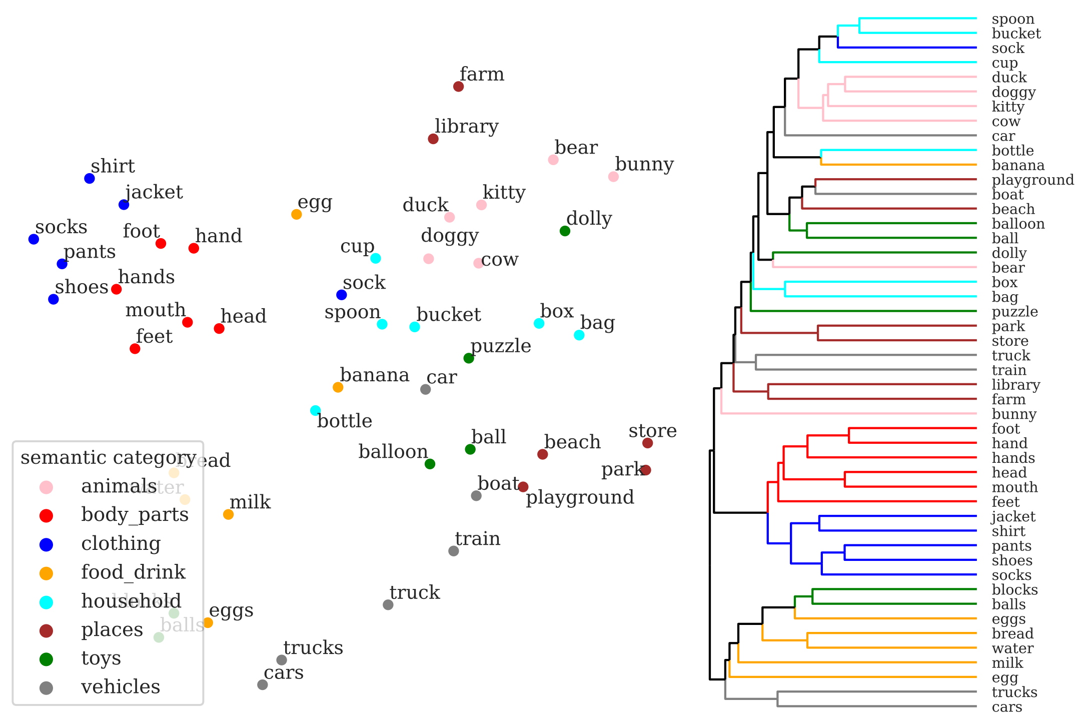
Additional cloze examples are shown in Table 5, showing the top model predictions for 3 runs of LSTM and CBOW. From these examples you can see networks are clearly forming word clusters corresponding to interpretable categories, not only larger syntactic categories like nouns and verbs, but also finer categories like animals and places, and other categories like be-verbs, words following "an", ditransitive verbs and V-ings. The LSTM tends to copy



(a) t-SNE.

(b) Dendrogram clustering.

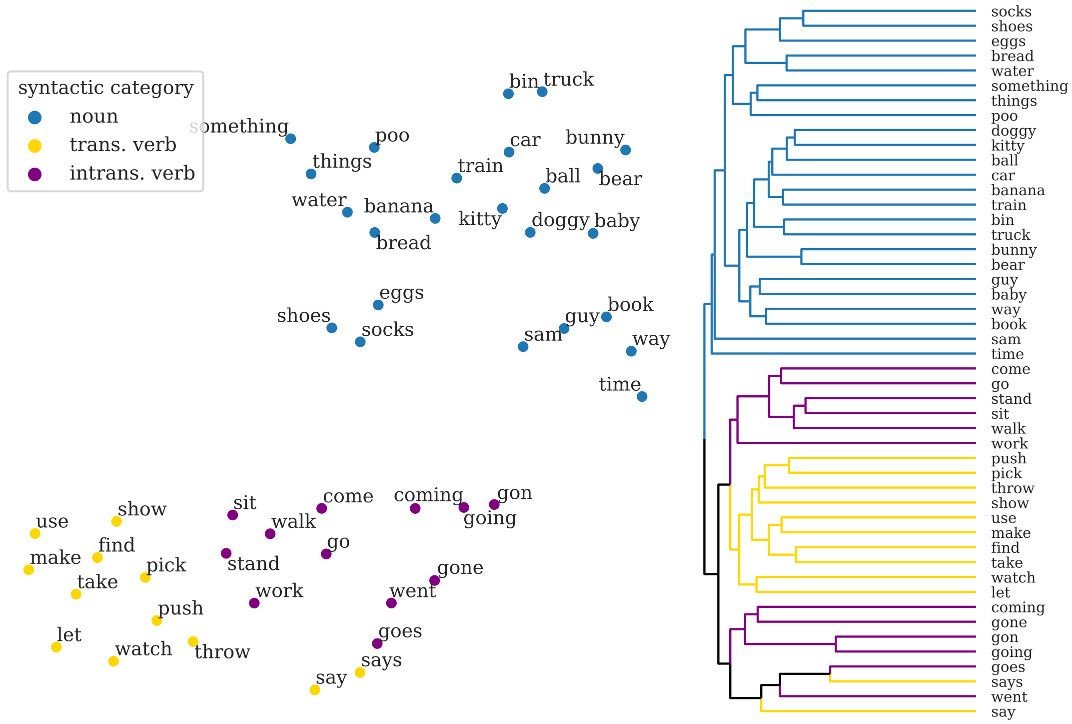
Figure 8: Clustering CBOW's word embeddings for syntactic categories by cosine measures in Figure 3. Nouns and verbs form two large clusters. Transitive and intransitive verbs form two smaller subclusters.



(a) t-SNE.

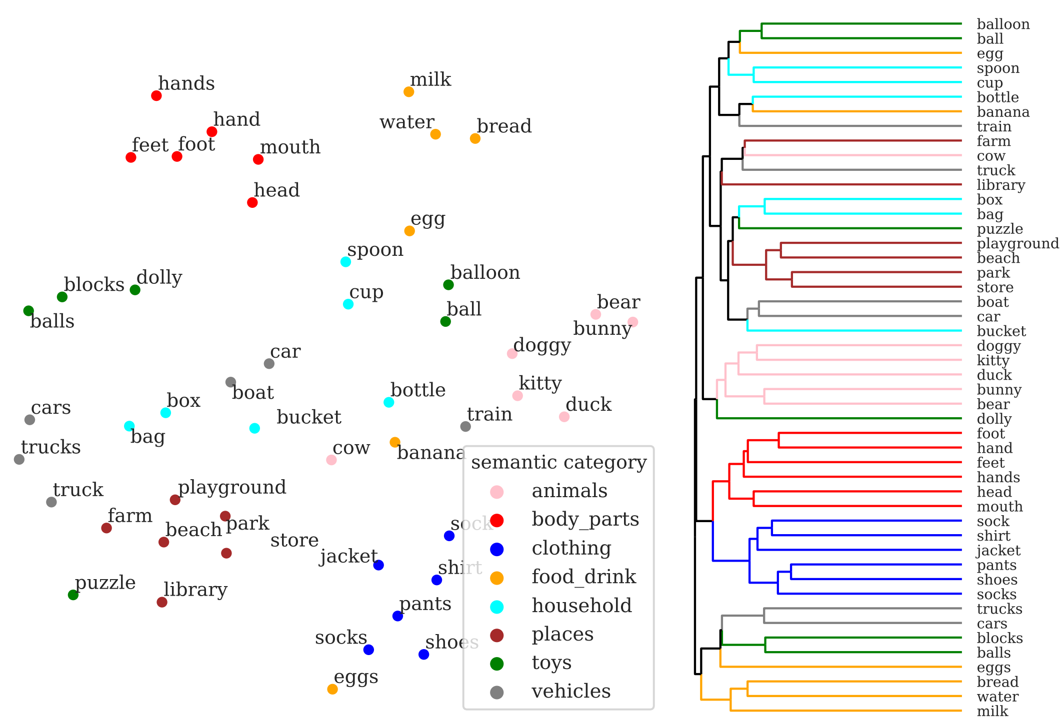
(b) Dendrogram clustering.

Figure 9: Clustering CBOW's word embeddings for semantic categories by cosine measures in Figure 3. The cluster structures are less clear than LSTM's in Figure 15, but several still correspond to semantic categories.



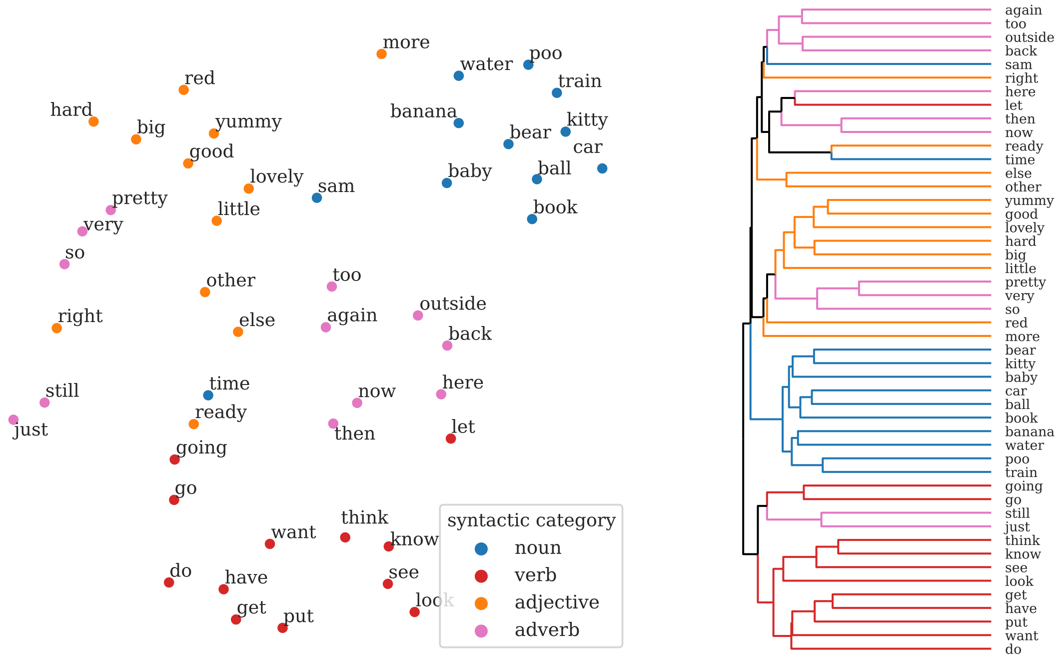
(a) t-SNE. (b) Dendrogram clustering.

Figure 10: Clustering Captioning LSTM's word embeddings for syntactic categories by cosine measures in Figure 3. Compared to Figure 3, the embedding structure is not significantly changed.



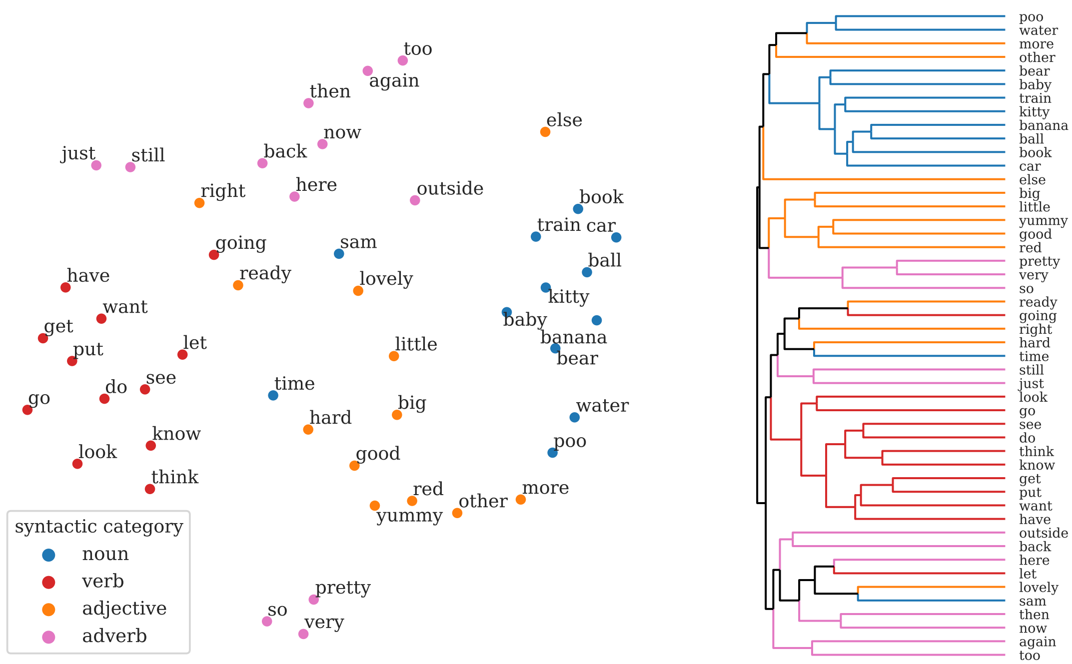
(a) t-SNE. (b) Dendrogram clustering.

Figure 11: Clustering Captioning LSTM's word embeddings for semantic categories by cosine measures in Figure 3. Compared to Figure 4, the embedding structure is not significantly changed.



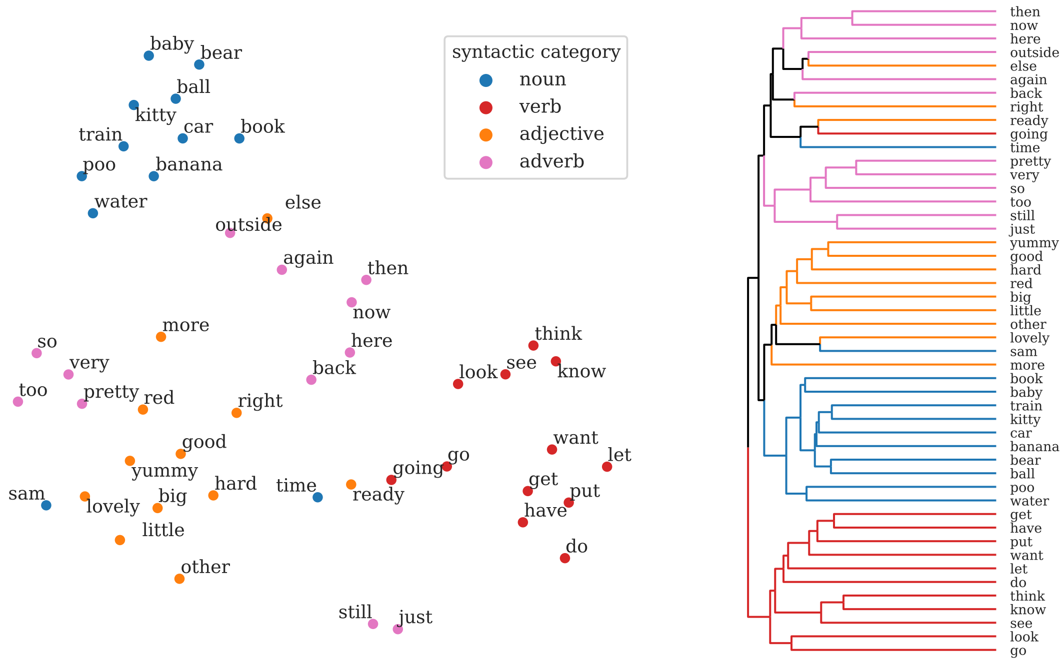
(a) t-SNE. (b) Dendrogram clustering.

Figure 12: Clustering LSTM's word embeddings for more syntactic categories by cosine measures in Figure 3. Clusters generally correspond to syntactic categories.



(a) t-SNE. (b) Dendrogram clustering.

Figure 13: Clustering CBOW's word embeddings for more syntactic categories by cosine measures in Figure 3. The cluster structures are also quite clear compared to LSTM's in Figure 12.



(a) t-SNE. (b) Dendrogram clustering.

Figure 14: Clustering Captioning LSTM's word embeddings for more syntactic categories by cosine measures in Figure 3. The cluster structures are also clear compared to those in Figure 12.

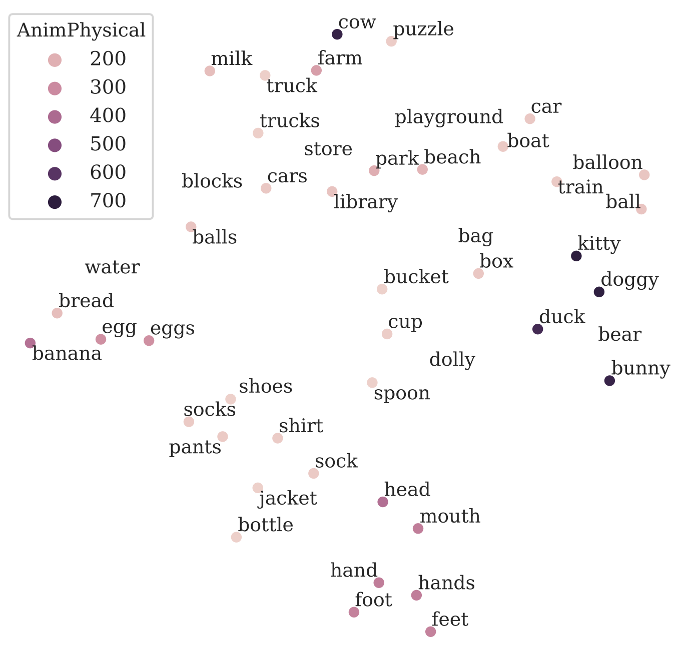


Figure 15: t-SNE of LSTM's word embeddings by distance $1-\cos(u, v)$. The set of words here is the same set in Figure 4. Hue means animacy. Our animacy data is from `https://osf.io/4t3cu/` contributed by Joshua VanArsdall and Janell Blunt. The animacy shown here is from their AnimPhysical field. For each word in our vocabulary, we try to get its animacy by looking up in the data its base form obtained by NLTK (Bird et al., 2009) lemmatizer; if not found, we do not include this word. Comparing this plot to Figure 4, we can see two clusters of animate categories at the bottom-right of the plot, corresponding to body parts and animals. This suggests that the embeddings in this plot may capture some animacy structure, although these results are closely aligned with semantic category structures due to the limited number of words shown in this plot.

Model	Top-5 predictions									
that's an o!										
LSTM	39.7%	egg	8.1%	emu	4.4%	eagle	4.2%	o	3.6%	ant
LSTM	44.9%	emu	6.6%	echidna	6.2%	ant	4.3%	egg	2.3%	s.
LSTM	25.2%	emu	14.3%	egg	9.5%	echidna	6.8%	ant	5.1%	o
CBOW	64.1%	hour	19.2%	ant	4.1%	apple	3.8%	emu	2.6%	egg
CBOW	53.3%	hour	16.7%	ant	8.1%	apple	6.6%	emu	5.3%	egg
CBOW	49.0%	hour	20.9%	ant	7.6%	apple	6.9%	emu	5.6%	egg
theres a strawberry and theres a flower										
LSTM	69.7%	s	17.7%	's	8.0%	is	1.8%	was	1.4%	's
LSTM	79.4%	s	10.6%	is	8.9%	's	0.4%	are	0.2%	was
LSTM	73.2%	s	16.7%	's	7.7%	is	0.9%	's	0.8%	was
CBOW	57.9%	's	20.7%	s	20.5%	is	0.5%	was	0.2%	are
CBOW	57.9%	's	20.8%	s	20.4%	is	0.4%	was	0.2%	are
CBOW	57.6%	's	21.0%	is	20.5%	s	0.4%	was	0.2%	are
theres a strawberry and theres a flower										
LSTM	26.6%	leaf	9.1%	car	9.0%	cardigan	7.0%	cupcake	5.4%	flower
LSTM	15.2%	ball	14.0%	strawberry	7.6%	bear	6.6%	banana	6.1%	kitty
LSTM	9.3%	cupcake	8.3%	kitty	5.5%	cup	5.3%	leaf	5.3%	cardigan
CBOW	9.2%	magazine	7.4%	bug	7.3%	moment	7.0%	biscuit	6.5%	horse
CBOW	10.0%	magazine	8.5%	bug	6.9%	moment	6.4%	biscuit	5.9%	horse
CBOW	9.7%	magazine	8.3%	moment	6.6%	bug	6.5%	biscuit	5.9%	horse
can you show me the eggs?										
LSTM	33.4%	give	25.9%	show	11.3%	tell	6.3%	pick	4.3%	get
LSTM	63.8%	show	21.2%	give	7.1%	get	1.7%	find	1.5%	throw
LSTM	56.2%	show	37.0%	give	1.8%	get	1.7%	throw	0.4%	lift
CBOW	61.5%	show	16.9%	give	11.2%	want	6.2%	tell	1.5%	showing
CBOW	62.3%	show	16.6%	give	11.2%	want	5.8%	tell	1.7%	showing
CBOW	63.2%	show	16.3%	give	11.0%	want	5.3%	tell	1.6%	showing
you keep eating.										
LSTM	30.2%	going	28.7%	trying	6.0%	eating	3.6%	done	2.8%	holding
LSTM	33.4%	going	11.2%	done	8.8%	eating	8.2%	trying	2.9%	doing
LSTM	24.2%	going	7.8%	looking	7.1%	doing	5.6%	eating	4.4%	one
CBOW	65.6%	going	14.4%	eating	4.6%	doing	4.6%	holding	2.4%	trying
CBOW	70.8%	going	7.9%	eating	5.6%	holding	3.5%	pressing	2.8%	trying
CBOW	69.2%	going	10.4%	eating	4.5%	trying	3.0%	doing	2.6%	pressing

Table 5: **Additional examples of clozes and the networks' predictions.** Three rows of a same architecture are results from three runs.

a word from the context, if that fits in the category. Also, the CBOW, which utilizes only near contexts, is doing surprisingly well, which indicates many unexpected correlations in the distributional patterns.

A.6 Linguistic Acceptability Analysis

As we mentioned in the main paper, we evaluated our networks on a subset of Zorro (Huebner et al., 2021), a minimal pair test suite consisting of 13 linguistic phenomena comprised of one or more subsets. Each subset contains 4,000 sentences making up 2,000 minimal pairs. Sentences in Zorro were created using templates filled with words from word lists they curated. Their word lists contained frequent words in the datasets they used. However, the word distribution in their datasets is different from ours. Among the 646 word types that occurred in Zorro, only 403 were in our vocabulary; most words not in our vocabulary were either human names, more abstract words usually not present in the early children's vocabulary (e.g., "control", "tradition", "bank"), or different word-forms (e.g., plural, past tense). Therefore, we filtered the sentence pairs so that they only consist of words contained within the vocabulary of our dataset. Table 6 lists the number of sentence pairs left in each subset, showing the remaining linguistic phenomena that we could evaluate our networks on.

The full set of results across each individual subset is shown in Figure 16. The Transformer network performs best on most of the tests. Note also that CBOW and the N-gram models do well on some, but not all subsets, and the LSTM is better overall. The N-gram models serve

Phenomenon	Subset	#sentence pairs left
agreement determiner noun	across 1 adjective	656
	between neighbors	616
agreement subject verb	across prepositional phrase	480
	across relative clause	532
	in question with aux	280
	in simple question	836
anaphor agreement	pronoun gender	0
argument structure	dropped argument	341
	swapped arguments	529
	transitive	384
binding	principle a	0
case	subjective pronoun	527
ellipsis	n-bar	0
filler-gap	wh-question object	0
	wh-question subject	0
irregular	verb	0
island-effects	adjunct island	0
	coordinate structure constraint	0
local attractor	in question with aux	480
npi licensing	matrix question	374
	only npi licensor	205
quantifiers	existential there	181
	superlative	188

Table 6: Number of sentence pairs left in each subset in Zorro. Each subset originally contained 2000 sentence pairs. After filtering, 15 out of 23 subsets, or 7 out of 13 phenomena, have sentence pairs left.

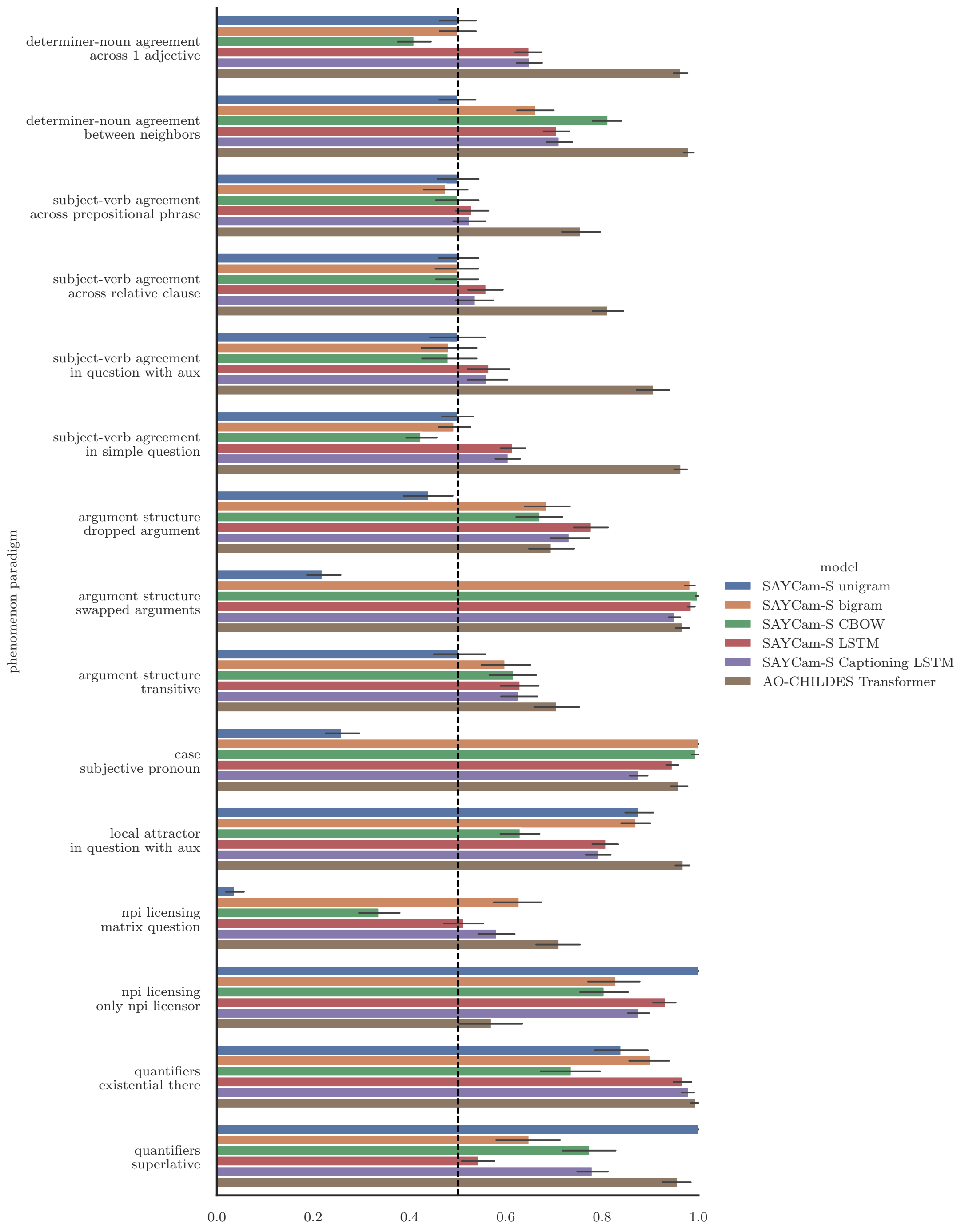


Figure 16: Accuracy on linguistic acceptability tests. The dashed line means the chance level.

Syntactic Category	#Types	Mean Loss Difference	t	p
all	617	-0.31	-11.42	<0.001
noun	220	-0.51	-9.44	<0.001
verb	150	-0.29	-5.58	<0.001
adjective	44	-0.14	-1.75	0.09
adverb	45	-0.14	-2.23	0.03
function word	82	-0.13	-3.25	0.002
cardinal number	11	-0.03	-0.18	0.86
.	65	-0.16	-2.09	0.04

Table 7: Type-level mean loss difference from language-only LSTM to Captioning LSTM on the validation set, with t-test results. Results on adjective and cardinal number are not significant.

as baselines indicating whether there are simple, short-distance or word count distributional cues in the data distribution that the model can potentially utilize. This turned out to be true with regard to some of the targeted tests. On the `quantifiers - superlative` subset, the unigram model achieves perfect accuracy because some quantifiers occur more frequently than others in the training data. For example, for superlative quantifiers "at least" and "more than", the product of unigram probabilities given the training corpus is higher for the latter which happens to be always grammatical in this subset. ("at": 682, "least": 11 vs. "more": 504, "than": 39). (Another pair of contrast in this subset, "at most" vs "fewer than", was filtered out because the word "fewer" is out-of-vocabulary.) Another example of short-distance distributional cues: in the `case - subjective pronoun` subset, the nominative case pronoun "I" usually occurs at the beginning of a grammatical sentence, so the bigram model always assigns a lower probability if it occurs not at the beginning of the sentence. The LSTM is better on subsets where longer-distance dependencies are required, such as `determiner-noun agreement - across 1 adjective` and `quantifiers - existential there`. The Captioning LSTM performs mostly close to the language-only LSTM; the only notable difference, shown in Figure 16, is that it is noticeably better on the `quantifiers - superlative` subset, close to the CBOW. We do not have a clear explanation for this performance difference, and more research is needed. Captioning cannot directly help in this task because the synthetic test sentences are not grounded and have no paired images; we simply fed the mean image of the training data (unrelated to the candidate sentences) to the captioning model when testing it on these candidate sentences. Though hypothetically captioning can indirectly help in training a stronger language model, due to the confound of the hidden state initialization in the captioning model and the specific data distribution of this subset.

A.7 Loss Difference between language-only LSTMs and Captioning LSTMs

Table 7 shows statistics of type-level loss difference between language-only LSTM to Captioning LSTM on the validation set, explaining Figure 6.

A.8 Cosine Similarity Heatmaps

Figures 17, 18 and 19 are heatmaps that visualize the cosine similarity matrices between words (corresponding to Figures 3, 10 and 8, respectively) for the LSTM, Captioning LSTM and CBOW models respectively, showing the similarity within and across different syntactic categories. These similarity matrices are used to calculate the Pearson correlations in the Representational Similarity Analysis in Section 4.2.2. Additionally, Figures 20, 21 and 22 are heatmaps that visualize the cosine similarity matrices between another set of words (corresponding to Figures 12, 14 and 13, respectively) for the LSTM, Captioning LSTM and CBOW models respectively, showing the similarity within and across another set of syntactic categories (noun, verb, adjective, adverb), from which we have the same observation as in Section 4.2.2 that LSTM and Captioning LSTM are very similar.

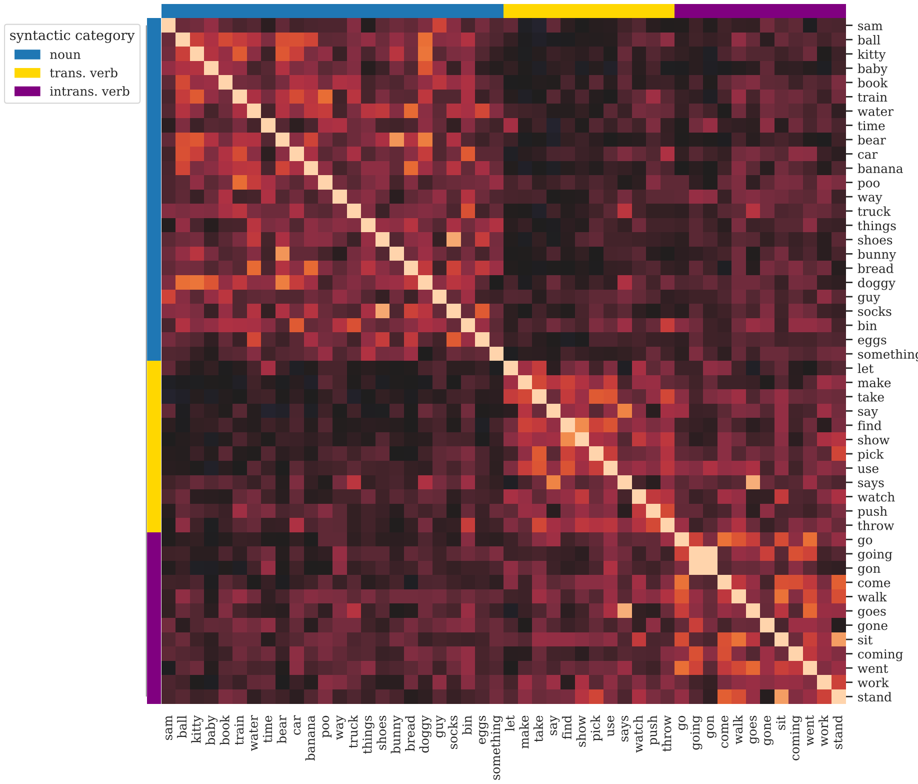


Figure 17: Heatmap of cosine similarity of LSTM's word embeddings. Nouns and verbs are more similar to other words within the same category than other words in the different category.

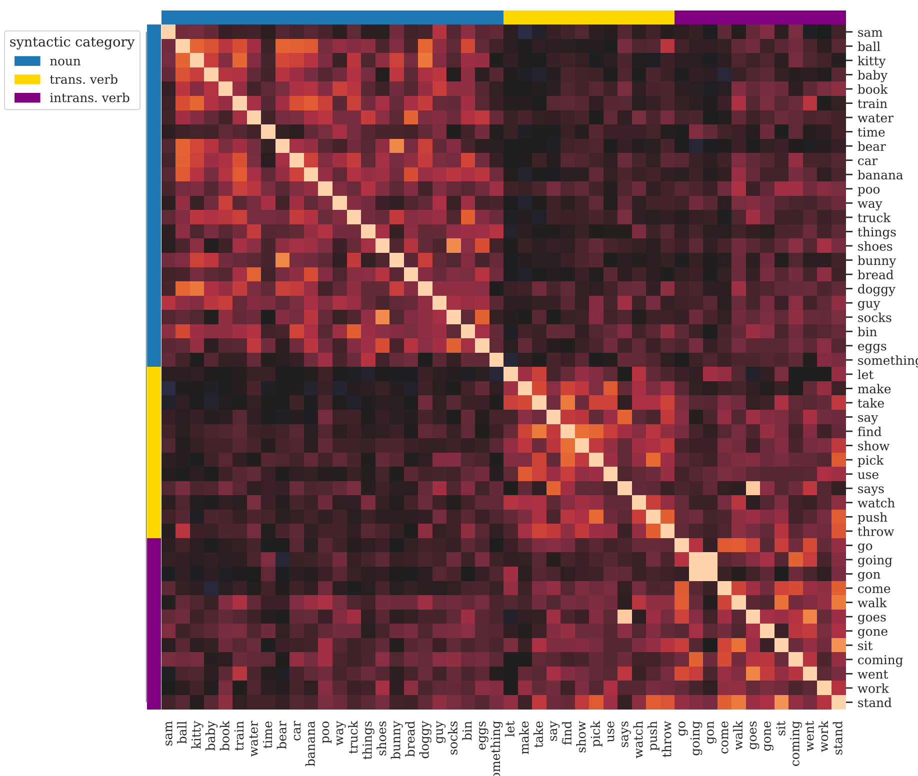


Figure 18: Heatmap of cosine similarity of Captioning LSTM's word embeddings. Nouns and verbs are more similar to other words within the same category than other words in the different category.

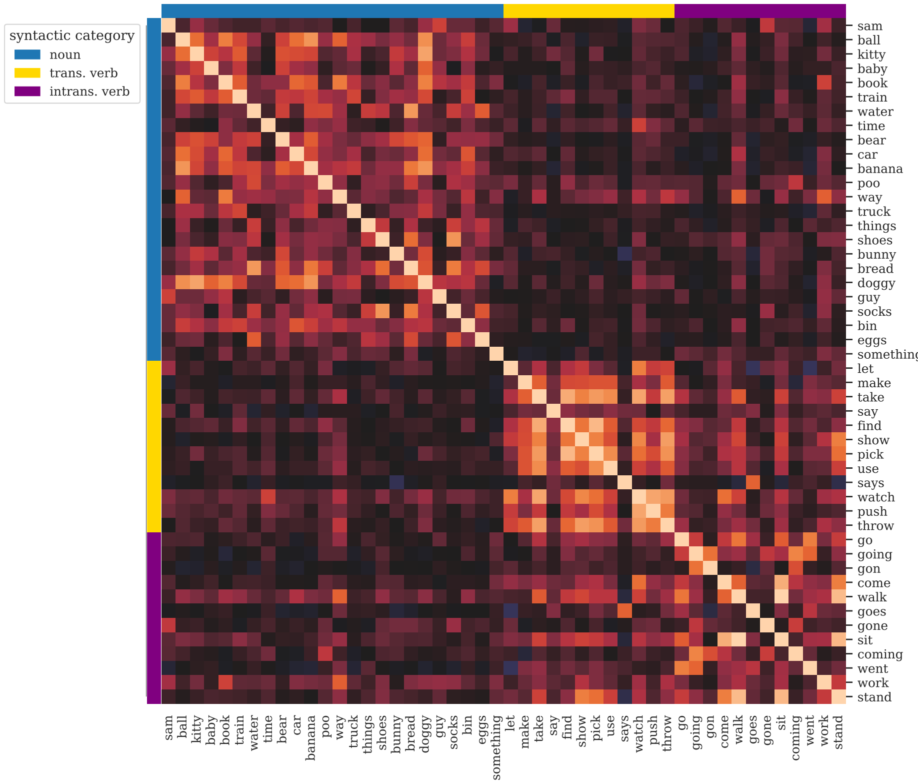


Figure 19: Heatmap of cosine similarity of CBOW's word embeddings. Nouns and verbs are more similar to other words within the same category than other words in the different category.

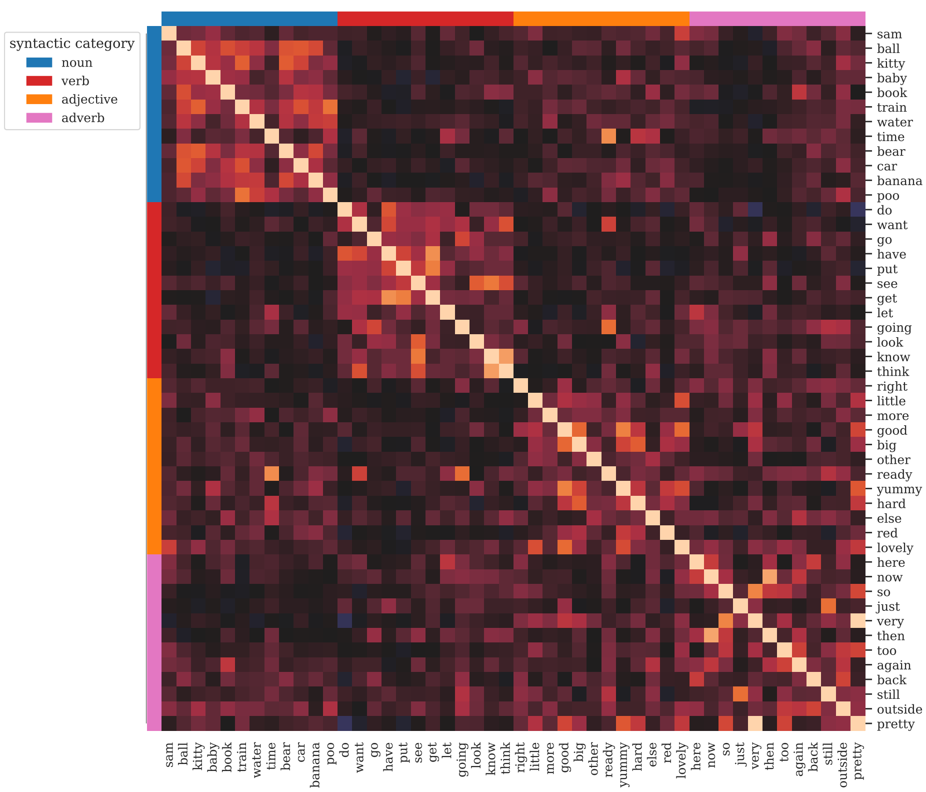


Figure 20: Heatmap of cosine similarity of LSTM's word embeddings. Words are more similar to other words within the same category than other words in the different category.

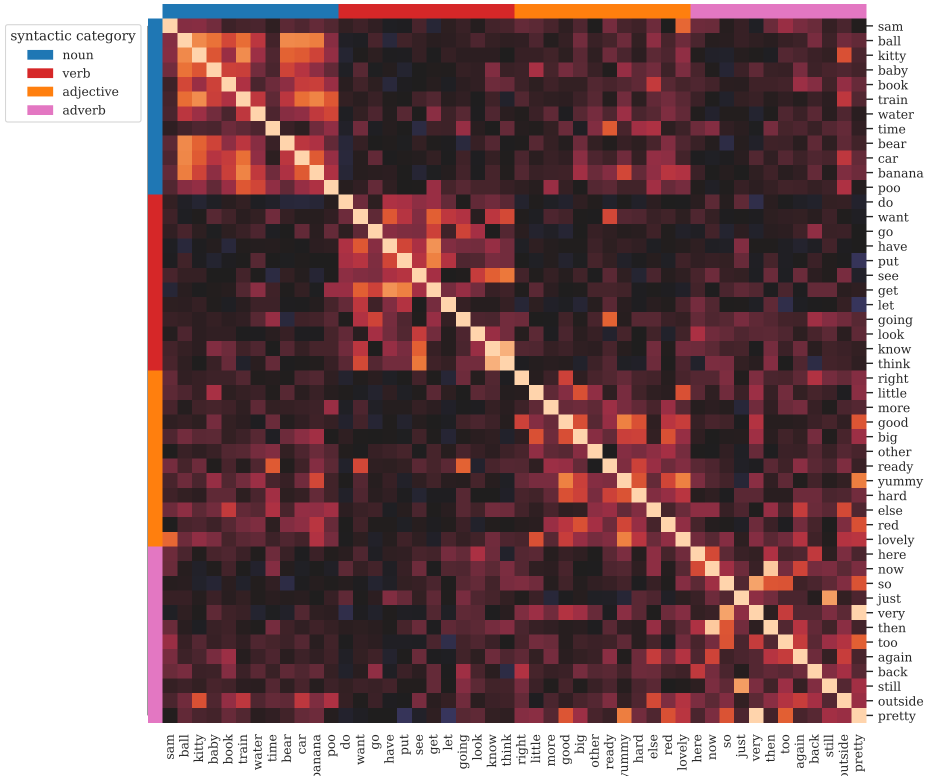


Figure 21: Heatmap of cosine similarity of Captioning LSTM's word embeddings. Words are more similar to other words within the same category than other words in the different category.

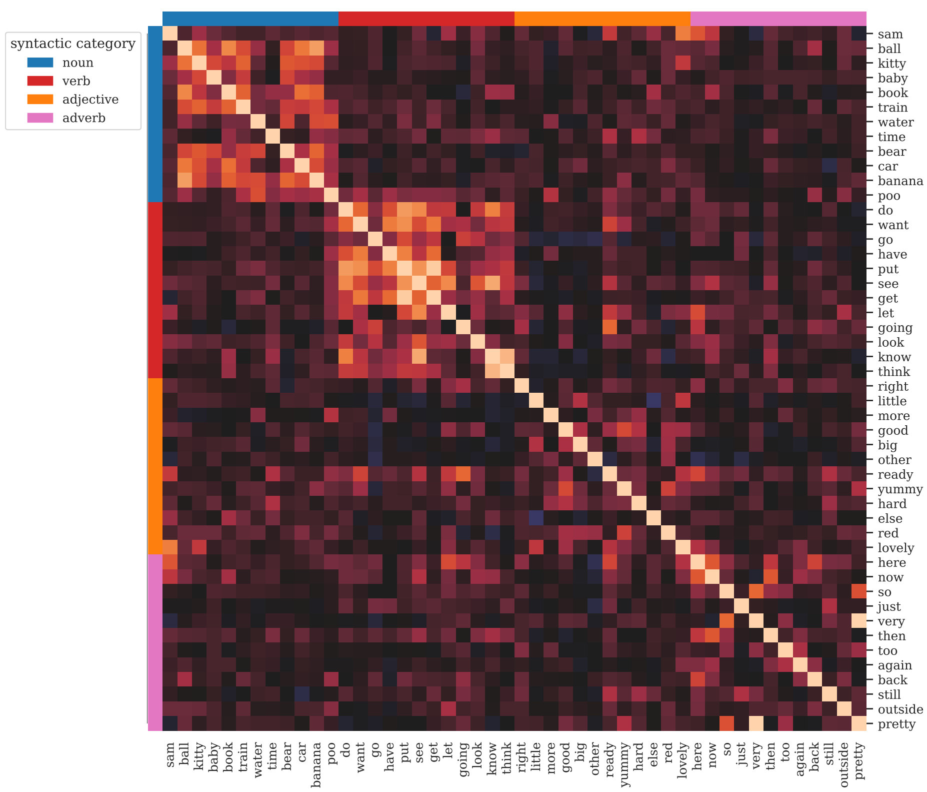


Figure 22: Heatmap of cosine similarity of CBOW's word embeddings. Words are more similar to other words within the same category than other words in the different category.