Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks

Mitja Nikolaus[1,2]
mitja.nikolaus@univ-amu.fr

Abdellah Fourtassi[1]
abdellah.fourtassi@gmail.com

[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2]Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France

Abstract

When learning their native language, children acquire the meanings of words and sentences from highly ambiguous input without much explicit supervision. One possible learning mechanism is cross-situational learning, which has been successfully tested in laboratory experiments with children. Here we use Artificial Neural Networks to test if this mechanism scales up to more natural language and visual scenes using a large dataset of crowd-sourced images with corresponding descriptions. We evaluate learning using a series of tasks inspired by methods commonly used in laboratory studies of language acquisition. We show that the model acquires rich semantic knowledge both at the word- and sentence-level, mirroring the patterns and trajectory of learning in early childhood. Our work highlights the usefulness of low-level co-occurrence statistics across modalities in facilitating the early acquisition of higher-level semantic knowledge.

1 Introduction

In order to acquire their native language, children learn both how to associate individual words with their meanings (e.g., the word "ball" refers to the object ball and the word "kick" refers to that act of kicking) and how to map the relationship between words in a sentence onto specific event configurations in the world, e.g., that the sequence of words "Jenny kicks the ball" maps on to the event where the referent of the first noun (i.e., Jenny) is performing the act of kicking on the second (i.e., the ball). This is a difficult task because it requires that children learn these associations and rules in a largely unsupervised fashion from an input that can be highly ambiguous (Quine, 1960). It is still unclear how children overcome this challenge.

Previous experimental studies on child language acquisition have focused on *evaluating* children's learning using controlled tasks that typically take the form of a two-alternative forced-choice paradigm. For example, in order to test the learning of an individual word meaning, we can utter this word to the child (e.g., "ball") and present her with two pictures representing correct (i.e., a ball) and incorrect referents (e.g. a cup), and we test if the child reliably prefers the correct one (Bergelson and Swingley, 2012). Similarly, in order to evaluate children's understanding of sentence-level semantics such as a the agent-patient relationship, we can utter a sentence such as "Jenny is tickling Mike" and present the child with two pictures where either Jenny or Mike are doing the tickling, and we test if the child reliably prefers the correct picture (e.g. Noble et al., 2011; Gertner and Fisher, 2012).

While we have been able to evaluate children's knowledge using such controlled tests, research has been less compelling regarding the *mechanism of learning* from the natural, ambiguous input. One promising proposal is that of cross-situational learning (hereafter, XSL). This proposal suggests that, even if one naming situation is highly ambiguous, being exposed to many situations allows the learner to narrow down, over time, the set of possible word-world associations (e.g. Pinker, 1989).

While in-lab work has shown that XSL is cognitively plausible using toy situations (Yu and Smith, 2007), effort is still ongoing to test if this mechanism scales up to more natural learning contexts using machine learning tools (e.g. Chrupała et al., 2015; Vong and Lake, 2020). This previous work, however, has focused mainly on testing the learning of individual words' meanings, while here we are interested in testing and comparing both word-level and sentence-level semantics.

1.1 The Current Study

The current study uses tools from Natural Language Processing (NLP) and computer vision as

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 200–210

research methods to advance our understanding of how unsupervised XSL could give rise to semantic knowledge. We aim at going beyond the limitations of in-lab XSL experiments with children (which have relied on too simplified learning input) while at the same time integrating the strength and precision of in-lab learning evaluation methods.

More precisely, we first design a model that learns in an XSL fashion from images and text based on a large-scale dataset of clipart images representing some real-life activities with corresponding – crowdsourced – descriptions. Second, we evaluate the model's learning on a subset of the data that we used to carefully design a series of controlled tasks inspired from methods used in laboratory testing with children. Crucially, we test the extent to which the model acquires various aspects of semantics both at the word level (e.g., the meanings of nouns, adjectives, and verbs) and at the sentence level (e.g. the semantic roles of the nouns).

Further, in order for an XSL-based model to provide a plausible language learning mechanism in early childhood, it should not only be able to succeed in the evaluation tasks, but also mirror children's learning trajectory (e.g., a bias to learn nouns before predicates). Thus, we record and analyze the model's learning trajectory by evaluating the learned semantics at multiple timesteps during the training phase.

1.2 Related Work and Novelty

While supervised learning from images and text has received much attention in the NLP and computer vision communities, for example in the form of classification problems (e.g. Yatskar et al., 2016) or question-answering (e.g. Antol et al., 2015; Hudson and Manning, 2019), here we focus on *cross-situational learning* of visually grounded semantics, which corresponds more to our understanding of how children learn language

There is a large body of work on cross-situational word learning (Frank et al., 2007; Yu and Ballard, 2007; Fazly et al., 2010), some of them with more plausible, naturalistic input in the form of images as we consider in our work (Kádár et al., 2015; Lazaridou et al., 2016; Vong and Lake, 2020). However, these previous studies only evaluate the semantics of single words in isolation (and sometimes only nouns). In contrast, our paper aims at a more comprehensive approach, testing and comparing the acquisition of both word-level meanings (including adjectives and verbs) and sentence-level semantics.

There has been some effort to test sentence-level semantics in a XLS settings. For example, Chrupała et al. (2015) also introduces a model that learns from a large-scale dataset of naturalistic images with corresponding texts. To evaluate sentence-level semantics, the model's performance was tested in a cross-modal retrieval task, as commonly used to evaluate image-sentence ranking models (Hodosh et al., 2013). They show that sentence to image retrieval accuracy decreases when using scrambled sentences, indicating that the model is sensitive to word order. In a subsequent study, Kádár et al. (2017) introduces *omission scores* to evaluate the models' selectivity to certain syntactic functions and lexical categories. Another evaluation method for sentence-level semantics is to compare learned sentence similarities to human similarity judgments (e.g. Merkx and Frank, 2019).

Nevertheless, these previous studies only explored broad relationships between sentences and pictures, they did not test the models' sensitivity to finer-grained phenomena such as dependencies between predicates (e.g., adjectives and verbs) and arguments (e.g., nouns) or semantic/ roles in detail.

2 Methods

2.1 Data

We used the Abstract Scenes dataset 1.1 (Zitnick and Parikh, 2013; Zitnick et al., 2013), which contains 10K crowd-sourced images each with 6 corresponding short descriptive captions in English. Annotators were asked to "create an illustration for a children's story book by creating a realistic scene" given a set of clip art objects (Zitnick and Parikh, 2013). The images contain one or two children engaged in different actions involving interactions with a set of objects and animals. Further, the children can have various emotional states depicted through a variety of facial expressions. The corresponding sentences were collected by asking annotators to write "simple sentences describing different parts of the scene"[1] (Zitnick et al., 2013).

While some studies have used larger datasets with more naturalistic images (e.g. Lin et al., 2014;

[1]The annotators were asked to refer to the children by the names "Jenny" and "Mike".

Plummer et al., 2015), here we used the Abstract Scenes dataset since it contains many similar scenes and sentences, allowing us to create balanced test sets (as described in the following section). In other words, the choice of the dataset was a trade-off between the naturalness of the images on the one hand and their partial systematicity, on the other hand, which we needed to design minimally different pairs of images to evaluate the model.

For the following experiments, we split the images and their corresponding descriptions into training (80%), validation (10%) and test set (10%).

2.2 Model

We use a modeling framework that instantiates XSL from images and texts in the dataset. To learn the alignment of visual and language representations, we employ an approach commonly used for the task of image-sentence ranking (Hodosh et al., 2013) and other multimodal XSL experiments (Chrupała et al., 2017; Vong et al., 2021).

The objective is to learn a joint multimodal embedding for the sentences and images, and to rank the images and sentences based on similarity in this space. State-of-the-art models extract image features from Convoluatinal Neural Networks (CNNs) and use LSTMs to generate sentence representations, both of which are projected into a joint embedding space using a linear transformation (Karpathy and Fei-Fei, 2015; Faghri et al., 2018).

As commonly applied in other multimodal XSL work (Chrupała et al., 2015; Khorrami and Räsänen, 2021), we assume that the visual system of the learner has already been developed to some degree and thus use a CNN pre-trained on ImageNet (Russakovsky et al., 2015) (but discard the final classification layer) to encode the images. Specifically, we use a ResNet 50[2] (He et al., 2016) to encode the images and train a linear embedding layer that maps the output of the pre-final layer of the CNN into the joint embedding space.

The words of a sentence are passed through a linear word embedding layer and then encoded using a one-layer LSTM (Hochreiter and Schmidhuber, 1997). Using a linear embedding layer, the hidden activations of the last timestep are then transformed into the joint embedding space.

[2]We also tried the more recent ResNet 152, but found results to be inferior. Also, we did not attempt to fine-tune the parameters of the CNN for the task, which could improve performance further.

The model is trained using a max-margin loss[3] which encourages aligned image-sentence pairs to have a higher similarity score than misaligned pairs, by a margin α:

$$\mathcal{L}(\theta) = \sum_a [\sum_b max(0, \gamma(i_a, s_b) - \gamma(i_a, s_a) + \alpha) + \sum_b max(0, \gamma(i_b, s_a) - \gamma(i_a, s_a) + \alpha)] \quad (1)$$

$\gamma(i_a, s_b)$ indicates the cosine similarity between an image i and a sentence s, (i_a, s_a) denotes a corresponding image-sentence pair. The loss is calculated for each mini-batch, negative examples are all examples in a mini-batch for which the sentence does not correspond to the image.

We train the model on the training set until the loss converges on the validation set. Details about hyperparameters can be found in the appendix.

2.3 Evaluation Method

In order to evaluate the model's acquisition of visually-grounded semantics, we used a two-alternative forced choice design, similar to what is typically done to evaluate children's knowledge in laboratory experiments (Bergelson and Swingley, 2012; Noble et al., 2011; Gertner and Fisher, 2012). Each test trial consists of an image, a target sentence and a distractor sentence: (i, s_t, s_d). We measure the model's accuracy at choosing the correct sentence given the image.

Crucially, we design the test tasks in a way that allows us to control for linguistic biases. Consider the example trial on the left in Figure 1. The model could posit that, say, Jenny (and not Mike) is the agent of an action even without considering the image, and only because Jenny may happen to be the agent in most sentences in the training data. To avoid such linguistic biases, we paired each test trial with a counter-balanced trial where the target and distractor sentence were flipped (cf. Figure 1, right side), in such a way that a language model without any visual grounding can only perform at chance level (50%).

[3]In preliminary experiments we also applied a max-margin loss with emphasis on hard negatives (Faghri et al., 2018), but observed a performance decrease. This could be due to the fact that our dataset contains many repeating sentences and semantically equivalent scenes, and consequently we could find "hard negatives" that should actually be positive learning examples (because they are semantically equivalent) in many situations.

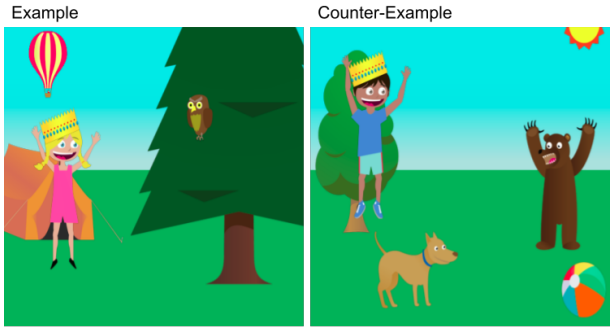

Figure 1: Counter-balanced evaluation of visually-grounded learning of semantics: Each test trial has a corresponding counter-example, where target and distractor sentence are flipped.

More precisely, we made the tasks as follows. First we searched in the heldout test set for image-sentence pairs $[(i_x, s_x), (i_y, s_y)]$ with *minimal differences* in the sentences given the phenomenon under study. For example, to study the acquisition of noun meanings, we look for pairs of sentences where the difference is only one noun such as $s_x =$ "jenny is wearing a crown" and $s_y =$ "mike is wearing a crown" (the corresponding images i_x and i_y depict the corresponding scenes, as shown in Figure 1). Second, based on such a minimal pair, we construct two counter-balanced triads: (i_x, s_x, s_y) and (i_y, s_y, s_x). The target sentence in one triad is the distractor in the other triad (and vice-versa). Using such a pair of counter-balanced triads, we test whether a model can both successfully choose the sentence mentioning "Jenny" when presented with the picture of Jenny *and* choose the sentence mentioning "Mike" when presented with the picture of Mike.

In the following we describe in more detail the phenomena of semantics we investigated using this testing setup. We provide an example for each category of task in Figure 2.

3 Tasks

3.1 Word-level Semantics

To study the acquisition of word meanings, we collect minimal pairs for the most commonly occurring nouns, adjectives and verbs. An example can be seen in Figure 1. Across all word-level categories, we make sure that there is only one referent present in the scene (this could be a child, an animal, or inanimate object, depending on the noun category under study). This ensures that we only evaluate word learning, and not more complex sentence-level semantics.[4]

Nouns We group the nouns into *persons*, *animals* and *objects*. Regarding persons, we consider the two children talked about in the dataset, i.e., *Jenny* and *Mike*. Regarding animals, we consider all 6 animals present in the dataset.[5] Regarding objects, we consider the 12 most frequently occurring words that are describing physical objects.[6]

Verbs The category of verbs is a bit tricky to evaluate because verbs are usually followed with an object that is tightly connected to them (e.g. *kicking* is usually connected to a ball whereas *eating* is connected to some food), resulting in a very limited availability of minimally different sentences with respect to verbs in the dataset. To be able to create a reasonable number of test trials, we trimmed the sentences[7] after the target verb and only consider verbs that can be used intransitively, e.g., "Mike is eating an apple" becomes "Mike is eating".

Further, we ensure, that the trials do not contain pairs of target and distractor sentences where the corresponding actions can be performed at the same time. For example, we do not include trials where the target sentence involves *sitting* and the distractor sentence *eating*, because the corresponding picture could be ambiguous: If the child in the picture is *sitting* and *eating* at the same, both the target and distractor sentences could be semantically correct. The resulting set of possible verb pairings is: ("sitting", "standing"), ("sitting", "running"), ("eating", "playing"), ("eating", "kicking"), ("throwing", "eating"), ("throwing", "kicking"), ("sitting", "kicking"), ("jumping", "sitting").

Adjectives The most common adjectives in the dataset are related to mood (e.g., happy and sad) and are displayed in the pictures using varied facial expressions (happy face vs sad face). Due to the lack of other kinds of adjectives[8], we only

[4] For example, if *Mike* (without a crown) was present in the picture to the left in Figure 1, the model would not only need to understand the difference between *Jenny* and *Mike*, but also understand what it means to *wear a crown* in order to correctly judge which sentence is the correct one, that is, which of Mike and Jenny is the one with the crown.

[5] ("dog", "cat", "snake", "bear", "duck", "owl")

[6] ("ball", "hat", "tree", "table", "sandbox", "slide", "sunglasses", "pie", "pizza", "hamburger", "balloons", "frisbee")

[7] The trimming was only done for the test trails and not in the training set.

[8] In the dataset, most of the properties for objects are fixed (e.g. colors and shapes) and are thus very rarely referred to in the descriptions. Consequently, we did not find minimal pairs

Figure 2: Examples for the evaluation of word and sentence-level semantics. Each test trial consists of an image, a target and a distractor sentence.

focused on mood-related adjectives. In addition, as there is no clear one-to-one mapping between each adjective and a facial expression, we only test the broad opposition between rather positive mood (smiling or laughing face) and rather negative mood (all other facial expressions). The resulting set of pairings was: ("happy", "sad"), ("happy", "angry"), ("happy", "upset"), ("happy", "scared"), ("happy", "mad"), ("happy", "afraid"), ("happy", "surprised").

Similar to what we did in the case of verbs, we trimmed the sentences after the target adjective in order to obtain more minimal pairs in our test set.

3.2 Sentence-level Semantics

In addition to evaluating the learning of word-level semantics, here we evaluate some (rudimentary) aspects of sentence-level semantics, that is, semantic phenomena where the model needs to leverage *relationships* between words in the sentence to be able to arrive at the correct solution. We focused on the following three cases for which a reasonable number of minimal pairs could be found.

Adjective - Noun Dependency In this task, we test if the model is capable of recognizing not only a given adjective (e.g., sad), but also the person experiencing this emotion (i.e. Jenny or Mike). The procedure used here is similar to the one we used to test individual adjectives, except that here the picture contains not only the person experiencing the target emotion but also the other person who is experiencing a different emotion (cf. examples on bottom left in Figure 2).

Take the following example: "mike is happy" and its minimally different distractor sentence "mike is sad" associated with a picture where Mike is happy and Jenny is sad (see Figure 2). In order to choose the target sentence over the distractor, the model needs to associate happiness with Mike but not with Jenny. In fact, since both persons appear in the picture and the word Mike appears in both sentences, the model cannot succeed by relying only on the individual name "mike" (in which case performance would be at chance). Similarly, it cannot succeed only by relying on the contrast "happy" vs. "sad" since Mike is happy but Jenny is sad (in which case performance would also be at chance).

Moreover, it cannot succeed even if it combines information in the words "mike" and "happiness" without taking into account their dependency in the sentence (say, if it only relied on a bag-of-

for adjectives describing simple properties like color.

	Evaluation task	Accuracy	p (best)	p (worst)	Size
Word-level Semantics	Nouns: Persons	0.78 ± 0.05	< 0.001	< 0.01	50
	Nouns: Animals	0.93 ± 0.02	< 0.001	< 0.001	360
	Nouns: Objects	0.86 ± 0.01	< 0.001	< 0.001	372
	Verbs	0.83 ± 0.05	< 0.001	< 0.001	77
	Adjectives	0.64 ± 0.06	< 0.01	0.25	56
Sentence-level Semantics	Adjective-noun dependencies	0.57 ± 0.01	< 0.05	< 0.05	192
	Verb-noun dependencies	0.72 ± 0.04	< 0.001	< 0.001	400
	Semantic roles	0.75 ± 0.06	< 0.001	< 0.05	50

Table 1: Accuracy, p-values (for the best and for the worst performing model) and evaluation set size (in number of trials) for all semantic evaluation tasks. The high variance in terms of number of trials is caused by the limited availability of appropriate examples in the dataset for some tasks (cf. Footnote 10).

words representation) because both the sentence and distractor would be technically correct in that case. More precisely, the bag of words of the target sentence {"mike", "happy"} and of the distractor {"mike", "sad"} both describe the scene accurately since the latter contains Mike, Happy, and Sad. The model can only succeed if it correctly learns that happiness is associated with Mike in the picture, suggesting that the model learns "happy" as modifier/predicate for "mike" in the sentence.

To construct test trials for this case, we used the same adjectives as for the word-level adjective learning, but we searched for minimal pair sentences with a second child in the scene with the opposite mood compared the target child.

Verb - Noun Dependencies Similar to adjective-noun dependencies, we aim to evaluate learning of verbs as predicate for the nouns they occur with in the sentence. We use the same verbs as in the word-learning setup as well as trim the sentences after the verb. We look for images with a target and distractor child engaged in different actions and construct our test dataset based on these scenes (see example in Figure 2, bottom right).

Semantic Roles In this evaluation, we test the model's learning of semantic roles in an action that involves two participants. We test the model's learning of the mapping of nouns to their semantic roles (e.g., agent vs. patient/recipient).

We look for scenes where both children are present and engaged in an action. In this action, one of the children is the agent and the other one is the patient/recipient. For example, in the sentence "jenny is waving to mike" the agent is Jenny and the recipient is Mike (see Figure 2, top right). The distractor sentence is constructed by flipping the subject and object in the sentence, i.e., "mike is waving to jenny". To succeed in the task, the model should be able to recognize that Jenny, not Mike, is the one doing the waving. This task is a more challenging version of the verb-noun dependency we described above because, here, Jenny and Mike are not only both present in the picture, they are also both mentioned in the sentences. To succeed, the model has to differentiate between agent and recipient in the sentence. Here again, a null hypothesis that assumes a bag-of-word representation of the sentence would not succeed: We need to take into account how each noun *relates* to the verb.

As with all other evaluation tasks, for each test trial we have a corresponding counter-balanced trial where the semantic roles are flipped.

4 Results

To evaluate the learned semantic knowledge, we measure, for each task, the model's accuracy at rating the similarity of the image and the target sentence $\gamma(i, s_t)$ higher than the similarity to the distractor sentence $\gamma(i, s_d)$. We report both final accuracy scores after the model has converged as well as intermediate scores before convergence, which we take as a proxy for the learning trajectory.

To ensure reproducibility, we make the semantic evaluation sets as well as the source code for all experiments publicly available.[9]

4.1 Acquisition Scores

We ran the model 5 times with different random initializations and evaluate each converged model

[9] https://github.com/mitjanikolaus/cross-situational-learning-abstract-scenes

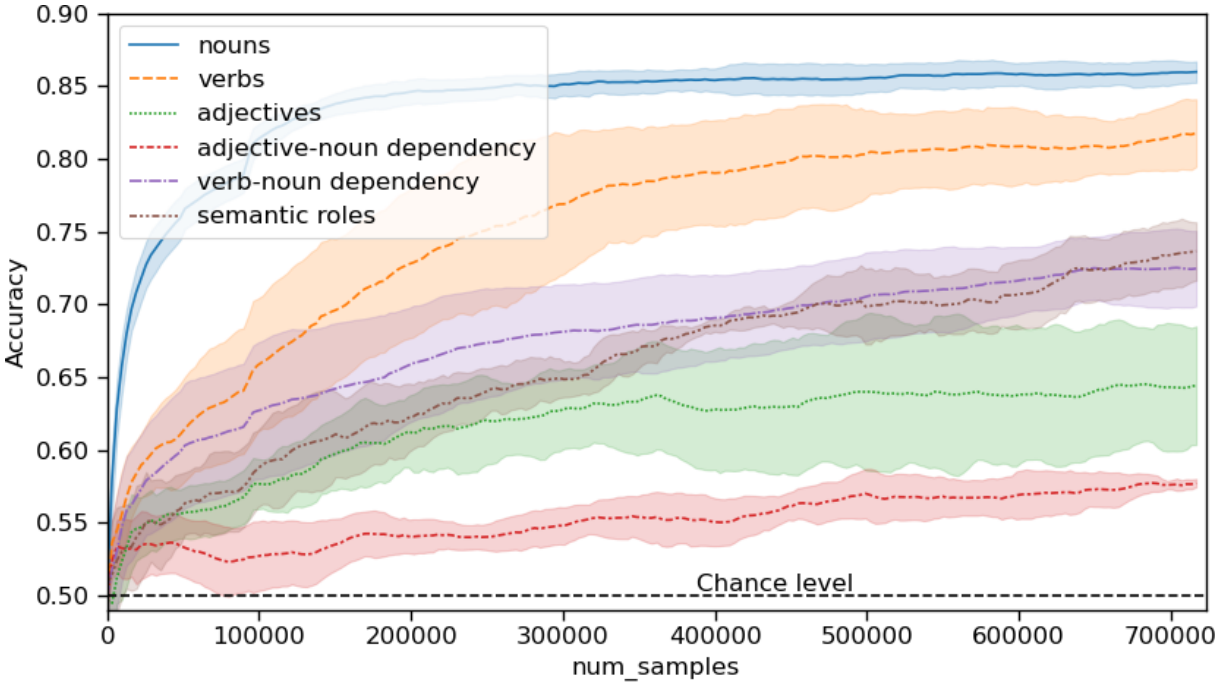

Figure 3: Learning trajectory of the models (mean over 5 runs, shaded areas show standard deviation). Accuracies for all noun categories were averaged. We calculated a rolling average over 30 data points to smooth the curve. The training set contains ~$50K$ examples, which means that the graph displays development over 15 epochs.

using the proposed tasks. Mean and standard deviation of the resulting accuracy scores can be found in Table 1. As some of the evaluation sets are rather small[10], we also performed binomial tests to evaluate whether the accuracy in the binary test is significantly above chance level (50%). We report the p-values' significance levels for the best and for the worst performing model[11] for each evaluation task.

The results show that the model has learned the semantics for most nouns very well. The score for verbs is also relatively high. As for adjectives, performance is only slightly above chance level and not always statistically significant, depending on the random initialization (e.g. the worst model is not significantly better than chance).

Regarding sentence-level semantics, the results suggest that the model has learned verb-noun dependencies and semantic roles relatively well. In contrast, Adjective-noun dependencies are not learned very well, which is not surprising given the poor adjective word-learning performance.

4.2 Acquisition Trajectories

In addition to the final evaluation scores, we are also interested in the *learning trajectory* of the model. We calculated the accuracy scores of the model every 100 batches. Figure 3 shows how the performance on the semantic evaluation tasks develops during the training of the model.

The model converged after having seen around 700K training examples (around 14 epochs). The trajectories show that the model first learns to discriminate nouns and only slightly later the verbs and then more complex sentence-level semantics.

5 Discussion

This paper dealt with the question of how children learn the word-world mapping in their native language. As a possible learning mechanism, we investigated XSL, that has received much attention in the literature. While laboratory studies on XSL have typically used very simplified learning situations to test if children are cognitively equipped to learn a toy language in an XSL fashion. The question remains as whether such a mechanism scales up to the learning of real languages where the learning situations can be highly ambiguous.

[10] Some evaluation sets are smaller than others due to the fact that all image-sentence pairs are taken directly from the test set and no new artificial images or sentences were created. This was done to ensure that the tests are performed using data that comes from the same distribution as the training set, i.e. data that the model has been exposed to.

[11] Each model corresponds to a different random initialization.

The novelty of our work is that we were interested not only in the scalability of XSL to learn from more naturalistic input, but also its scalability to the learning of various aspects of semantic knowledge. These include both the meanings of individual words (belonging to various categories such as nouns, adjectives, and verbs) and the meanings of higher level semantics such as the ability to map how words relate to each other in the sentence (e.g., subject vs. object) to the semantic roles of their respective referent in the world (e.g., agent vs. patient/recipient). We were able to perform these evaluations using a simple method inspired from the field of experimental child development and which has usually been used to test the same learning phenomena in children, i.e., the two-alternative forced choice task.

Using this evaluation method, we found that an XSL-based model trained on a large set of pictures and their descriptions was able to learn word-level meanings for nouns and verbs relatively well, but struggles with adjectives. Further, the model seems to learn some sentence-level semantics, especially verb-noun dependencies and semantic roles. Finally, concerning the learning trajectory, the model initially learns the semantics of nouns and only later the semantics of verbs and more complex sentence-level semantics.

Concerning word-level semantics, the fact that the model learns nouns better than (and before) the predicates (adjectives and verbs) resonates with findings in child development about the "noun bias" (Gentner, 1982; Bates et al., 1994; Frank et al., 2021). The model also learns verbs better than adjectives. However, we suspect this finding is caused by the limited availability of adjectives in the dataset.[12] In fact, the verb-related actions (e.g. "sitting" vs. "standing") were arguably more salient and easier to detect visually than adjective-related words ("happy" vs. "sad") which require a fine-grained detection of the facial expressions.

Concerning sentence-level semantics, the model performed surprisingly well on verb-noun dependency task where the model assigned a semantic role to one participant and on the similar but (arguably) more challenging task of assigning semantic roles to two participants. Further, the fact that the model shows a rather late onset of understanding of semantic roles, only after a set of nouns and verbs have been acquired (cf. Figure 3) mirrors children's developmental timeline. Indeed, children become able to assign semantic roles to nouns in a sentence correctly when they are around 2 years and 3 months old (Noble et al., 2011), at an age when they have already acquired a substantial vocabulary including many lexical categories such as nouns and verbs (Frank et al., 2021)

In this paper, we used artificial neural networks to study how properties the input can (ideally) inform the learning of semantics. Our modeling did not purport to account for the details of the cognitive processes that operate in children's minds nor did it take into account limitations in children's information-processing abilities. Thus, this work is best situated at the computational level of analysis (Marr, 1982), which is only a first step towards a deeper understanding of the precise algorithmic implementation. That said, we can speculate about the internal mechanisms used by the model to succeed in the tasks and about their potential insights into children's own learning. For example, it is very likely that the model leverages simple heuristics to recognize the agent in a sentence, e.g., it may have learned to associate the first appearing noun in the sentence to the agent of the action. Research on child language suggest that children also use such heuristics (e.g. Gertner and Fisher, 2012). This suggests that the model, like children, might use partial representations of sentence structure (i.e., rudimentary syntax) to guide semantic interpretation.

Exploiting structural properties of the input (e.g., order of words in a sentence) may be insightful when it mirrors genuine learning heuristics in children. However, a neural network model may also capitalize on idiosyncratic biases in the dataset (that do not reflect the natural distribution in the world) to achieve misleadingly high performance.[13] For example, a misleading bias in the linguistic input is if a certain noun (e.g., Jenny) occurs more frequently in the dataset as agent, leading the model to, say, systematically map "Jenny" to agent. Similarly, an example of a misleading bias in visual data is if the agent is always depicted on the left or right side of the image, leading the model to capitalize on this artificial shortcut.

In the current work, we controlled for linguistic biases by counter-balancing all testing trials. As for the visual bias, we ruled out some artificial bi-

[12] The data contained mostly mood-related adjectives.

[13] For example, Goyal et al. (2017) finds that grounded language models trained on a visual question answering task are exploiting linguistic biases of the training set.

ases such as the agent spatial order in the images. Indeed, investigation of our semantic roles test set shows that the agent occurs roughly equally on the right (52%) and left sides, which means that a model exploiting such a bias could only perform around chance level. There could be other biases we are not aware of and which require performing further controls. That said, this is an open question for all research using neural networks as models of human learning. More generally, our understanding of language acquisition would greatly benefit from further research on the interpretation of neural network learning, revealing the content of these black box models. This would allow us to tease apart genuine insights about realistic heuristics that could be used by children and artificial shortcuts that only reflect biases in the learning datasets.

In future work, we plan to study visual datasets with even more naturalistic scenes such as COCO (Lin et al., 2014). In this regard, maybe closer to our work is the study by Shekhar et al. (2017a,b) who used COCO to create a set of distractor captions to analyze whether vision and language models are sensitive to (maximally difficult) single-word replacements. Our goal is to go beyond these analysis to test specific semantic phenomena as we did here with the Abstract Scenes dataset. Another step towards more naturalistic input is the use speech input instead of text (Chrupała et al., 2017; Khorrami and Räsänen, 2021).

Finally, this work focused on testing how XSL scales up to natural language learning across many semantic tasks. Nevertheless, children's language learning involves more than the mere tracking of co-occurrence statistics: They are also social beings, they actively interact with more knowledgeable people around them and are able to learn from such interactions (Tomasello, 2010). Future modeling work should seek to integrate both statistical and social learning skills for a better understanding of early language learning.

Acknowledgements

We thank Mostafa Abdou, Jasper Bischofberger, Bissera Ivanova, Chiara Mazzocconi and the reviewers for their feedback and comments.

This work, carried out within the Labex BLRI (ANR-11-LABX-0036) and the Institut Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government, managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX)

A Appendix

A.1 Model Details

The hyperparameters of the model were chosen on general best-practices and not any further tuned.

Minimum word frequency for vocab	5
Word Embeddings Size	100
Joint Embeddings Size	512
LSTM Hidden Layer Size	512
Optimizer	Adam
Initial Learning Rate	0.0001
Batch size	32
α (margin for loss term)	0.2

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Elizabeth Bates, Virginia Marchman, Donna Thal, Larry Fenson, Philip Dale, J Steven Reznick, Judy Reilly, and Jeff Hartung. 1994. Developmental and stylistic variation in the composition of early vocabulary. *Journal of child language*, 21(1):85–123.

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

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 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.

Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 112–118, Beijing, China. Association for Computational Linguistics.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2021. *Variability and consistency in early language learning: The Wordbank project*. MIT Press.

Michael C Frank, Noah D Goodman, and Joshua B Tenebaum. 2007. A bayesian framework for cross-situational word-learning.

Dedre Gentner. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*.

Yael Gertner and Cynthia Fisher. 2012. Predicted errors in children's early sentence comprehension. *Cognition*, 124(1):85–94.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Ákos Kádár, Afra Alishahi, and Grzegorz Chrupała. 2015. Learning word meanings from images of natural scenes. *Traitement Automatique des Langues*, 55(3).

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Khazar Khorrami and Okko Räsänen. 2021. Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - a computational investigation.

Angeliki Lazaridou, Grzegorz Chrupała, Raquel Fernández, and Marco Baroni. 2016. Multimodal semantic learning from child-directed input. In *Knight K, Nenkova A, Rambow O, editors. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12-17; San Diegio, California. Stroudsburg (PA): Association for Computational Linguistics; 2016. p. 387–92*. ACL (Association for Computational Linguistics).

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

David Marr. 1982. Vision: A computational investigation into the human representation and processing of visual information.

Danny Merkx and Stefan L Frank. 2019. Learning semantic sentence representations from visually grounded language without lexical knowledge. *Natural Language Engineering*, 25(4):451–466.

Claire H Noble, Caroline F Rowland, and Julian M Pine. 2011. Comprehension of argument structure and semantic roles: Evidence from english-learning children and the forced-choice pointing paradigm. *Cognitive science*, 35(5):963–982.

Steven Pinker. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT press.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Willard Van Orman Quine. 1960. *Word and object*. MIT Press.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. Vision and language integration: Moving beyond objects. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265.

Michael Tomasello. 2010. *Origins of human communication*. MIT press.

Wai Keen Vong and Brenden M. Lake. 2020. Learning word-referent mappings and concepts from raw inputs. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020*. cognitivesciencesociety.org.

Wai Keen Vong, Emin Orhan, and Brenden Lake. 2021. Cross-situational word learning from naturalistic headcam data. In *34th CUNY Conference on Human Sentence Processing*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542.

Chen Yu and Dana H Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.

Chen Yu and Linda B Smith. 2007. Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, 18(5):414–420.

C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016.

C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. 2013. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1681–1688.