Modeling the Interaction Between Perception-Based and Production-Based Learning in Children's Early Acquisition of Semantic Knowledge

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

Abdellah Fourtassi[1]
abdellah.fourtassi@univ-amu.fr

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

Abstract

Children learn the meaning of words and sentences in their native language at an impressive speed and from highly ambiguous input. To account for this learning, previous computational modeling has focused mainly on the study of perception-based mechanisms like cross-situational learning. However, children do not learn only by exposure to the input. As soon as they start to talk, they practice their knowledge in social interactions and they receive feedback from their caregivers. In this work, we propose a model integrating both perception- and production-based learning using artificial neural networks which we train on a large corpus of crowd-sourced images with corresponding descriptions. We found that production-based learning improves performance above and beyond perception-based learning across a wide range of semantic tasks including both word- and sentence-level semantics. In addition, we documented a synergy between these two mechanisms, where their alternation allows the model to converge on more balanced semantic knowledge. The broader impact of this work is to highlight the importance of modeling language learning in the context of social interactions where children are not only understood as passively absorbing the input, but also as actively participating in the construction of their linguistic knowledge.

1 Introduction

An important aspect of language acquisition is learning how to map linguistic forms to meanings. This involves both mapping individual word forms (e.g., "dog") to concepts of the world (e.g. the category DOG) and mapping the relationship between words in a sentence (e.g., "the dog chases the ball") to a given event configuration in the world (i.e., that the dog is the agent performing the act of chasing on the ball, the semantic patient). Children manage to learn this mapping in their native language at an impressive speed (Fisher and Gleitman, 2002; Golinkoff et al., 2013; Frank et al., 2021) and despite the high ambiguity of this task in the natural context where language learning occurs (Quine, 1960).

Perception-based learning

Much modeling effort has focused on learning from the multimodal input that children *perceive* around them. These models are based on Cross-Situational Learning (hereafter XSL): While a single word-world mapping situation is ambiguous, being exposed to many situations allows the learner to narrow down, over time, the set of possible associations. This kind of learning has been demonstrated using toy situations in controlled laboratory testing with children (Smith and Yu, 2008). It has also been shown to scale up to more realistic learning contexts using a combination of NLP and computer vision tools (Chrupała et al., 2015; Vong and Lake, 2020, 2021).

Production-based learning

Learning from perceived multimodal input is an important mechanism, especially in the early stages of development. Nevertheless, an additional mechanism comes into play as soon as children start to *produce* language themselves, thus becoming able to receive *feedback* from more linguistically knowledgeable interlocutors (e.g., caregivers) (Warlaumont et al., 2014; Clark, 2018; Tsuji et al., 2021).

One specific form of feedback that has received much attention is when the caregiver provides explicit reformulation to the child's inadequate use of words (Brown, 1973; Chouinard and Clark, 2003; Saxton et al., 2005; Hiller and Fernández, 2016). Nevertheless, explicit reformulation is not the only

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 391–407
November 10–11, 2021. ©2021 Association for Computational Linguistics

way children can get useful feedback on their early production. For instance, the feedback that signals communicative success/failure to the child – even in an implicit form – can also play a role. Below we elaborate on the nature and potential usefulness of this – more general – mechanism which we call **Communicative Feedback** (hereafter CF).

When children start to talk, they immediately start putting words to use in social interaction to try and establish coordinated communication. This coordination aims at achieving various goals such as directing the interlocutor's attention (e.g., "A duck!") or requesting something (e.g., "I am thirsty!"), among many other communicative intents that children demonstrate very early in life (Snow et al., 1996; Casillas and Hilbrink, 2020; Nikolaus et al., 2021).

Importantly, children are sensitive to when coordination appears to *break down* without necessarily requiring explicit correction or even a verbal response from the caregiver. In fact, the child might feel misunderstood merely by not getting the reaction she expected (e.g., a puzzled or a still face) or by not getting the exact object she requested (e.g. Tronick et al., 1978; Markova and Legerstee, 2006). On such occasions, the child may not be offered the correct linguistic form as in reformulation-based (or corrective) feedback, but communication breakdown represents in and of itself a negative feedback, a cue to the child that her way of using words was not correct and that it should be revised for communication to be re-established or "repaired" (Clark, 2018, 2020). Vice versa, successful coordination (i.e., a contingent response or action from the caregiver) signals to the child that her use of words was probably adequate, encouraging (or reinforcing) this use in future conversations.

Compared to explicit corrective feedback, CF relies on sensitivity to broad coordination and miscoordination cues that are fundamental to reach shared understanding in any linguistic exchange (see "communicative grounding" (Clark, 1996)). It is, thus, arguably more pervasive in child-caregiver conversations and less dependent on parenting styles, Socioeconomic Status (SES) or culture (Childers et al., 2007; Mesman et al., 2018).

Previous experimental research has explored a simple form of Communicative Feedback and how it can help with language acquisition, especially regarding the emergence of speech-related vocalization (Oller, 2000). When the child produces a sound that contains speech-related vocalization (as opposed to other non-speech types of vocalization such as cry or laugh), the child is more likely to receive an immediate, positive response from the caregiver than if the produced sound is not speech-related. Critically, the fact of receiving a response from the caregiver (that is contingent of the production of speech) encourages the child to subsequently produce more speech-related vocalizations (Bloom, 1988; Goldstein et al., 2003; Goldstein and Schwade, 2008; Warlaumont et al., 2014).

To the best of our knowledge, no previous modeling work has investigated the role that CF could play in *semantic learning* or how CF may interact with the – more studied – class of semantic learning mechanisms that are based on perception alone such as XSL.

1.1 The current study

This work aims at providing a comprehensive account of early semantic learning combining both perception-based learning through XSL and production-based learning through CF. The learning account we propose is very similar to – and in fact, can be seen as a computational instantiation of – the "original word game" proposed by Brown (1958):

> "The original word game is the operation of linguistic reference in first language learning. At least two people are required: One who knows the language (the tutor) and one who is learning (the player) ... The tutor names things in accordance with the semantic customs of the community. The player forms hypotheses about the categorical nature of the things named. He tests his hypotheses by trying to name new things correctly. The tutor compares the player's utterances with his own anticipations of such utterances and, in this way, checks the accuracy of fit between his own categories and those of the player. He improves the fit by correction."[1]

Here we focus on a simple case of semantic learning where the meaning can be derived from

[1] Note that in our work, the fit of the semantic knowledge is not necessarily improved by *correction*, but rather by broad cues about the success or failure of the communicative coordination.

concrete visual scenes. We use an integrated model to characterize the child's learning both during the perception phase and during the production phase. In the perception phase, the model optimizes the generative probability of the tutor's utterances given the visual scenes. This probability is refined thanks to exposure to several situations (i.e., XSL).

In the production phase, the same language model is now used to generate utterances given a scene. The adequacy of the utterance is evaluated against the gold standard descriptions of the scene (representing the tutor's superior knowledge). The adequacy value is a continuous number we use to characterize the valence of the Communicative Feedback: The higher the adequacy, the more likely the child receives signals of communication success from the tutor (e.g., a positive, contingent reaction). Vice versa, the lower the value, the more likely the child receives signals of communication breakdown (e.g. a puzzled face or a non-contingent reaction). The model gets updated via Reinforcement Learning (RL) using the adequacy value as a reward.

Using this computational framework, we study the role of CF in early semantics acquisition. In addition, we investigate how CF interacts with XSL. We evaluate and compare these two mechanisms in terms of how they fare on a wide range of semantic tasks including both word-level (nouns, adjectives, and verbs) and sentence-level meaning acquisition (e.g. semantic roles).

Combining some kind of (weakly) supervised learning model with reinforcement learning is not a new technique. Such a setup has been used in previous NLP work (Ranzato et al., 2016; Rennie et al., 2017). The novelty of our work is to use these tools to instantiate new hypotheses about early language acquisition and to test these hypotheses using a benchmark of language acquisition tasks, similar to the tasks used to study children's semantic learning in laboratory experiment.

The paper is organized as follows. First we present the cross-modal dataset we use in this work and introduce the modeling framework. We explain how we instantiate both the perception-based mechanism (XSL) and the production-based mechanism (CF) using tools from NLP and computer vision. Next, we present the experiments we run: each representing a learning scenario, including scenarios combining both perception and production-based mechanisms. Next, we test the extent to which these models learn various aspects of semantics. Finally, we discuss the results in the light of the literature on early language learning.

To ensure reproducibility, we make the source code for the model and all experiments publicly available.[2]

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. The images are clip-art scenes involving one or two children engaged in different actions involving a set of different objects and animals.[3] The corresponding captions were crowd-sourced from a different set of annotators.[4] Two example scenes along with descriptions can be found in Figure 1.

We use this dataset as it allows us to evaluate the learning of visually-grounded semantics on the word-level and sentence-level, using recently proposed evaluation tasks by (Nikolaus and Fourtassi, 2021) (see also Section 2.4). Other studies on XSL have used larger dataset with naturalistic images (e.g. Lin et al., 2014; Plummer et al., 2015). However, there is currently no similar evaluation method available for these datasets that allows for detailed examination of the learned visually grounded semantics. We divide the data into training (80%), validation (10%) and test splits (10%) as proposed in Nikolaus and Fourtassi (2021).

2.2 Modeling framework

We develop an integrated modeling framework that can both learn from pairs of images and sentences in the context of XSL *and* and to produce its own sentences given an image to learn using rewards (CF). This framework will allow to assess various learning scenario, including ones that combine both XSL and CF.

Some previous work in NLP has used image-sentence ranking models (Hodosh et al., 2013) to

[2] https://github.com/mitjanikolaus/perception-and-production-based-learning

[3] 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).

[4] Annotators were asked to write "simple sentences describing different parts of the scene". They were asked to refer to the children by the names "Jenny" and "Mike" (Zitnick et al., 2013).

learn the alignment of visual and language representations, and thus to model cross-modal XSL (Chrupała et al., 2017; Vong et al., 2021; Nikolaus and Fourtassi, 2021). However, these models are not designed to *produce* new utterances given an image.

As we are here interested in both perception and production, we use a different computational framework borrowed from studies on image captioning (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018). This framework is based on a language model conditioned on the image. Just like the image-sentence ranking models, here the model is trained using pairs of images and captions, instantiating learning in a XSL fashion. In addition, the same language model can be used to generate sentences given an image, which we used to instantiate the production-based mechanism CF. Since the goal is not to produce a state-of-art image captioning model, we consider a basic implementation close to that used in Vinyals et al. (2015).

To process the images, we use ResNet 50 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015), assuming that the visual system of the child has already been developed to some degree allowing her to process visual scene.[5] We discard the final classification layer and fine-tune the remaining layers of this CNN during the training progress to encode the images in our dataset.

Conditioned on this image encoding, an autoregressive language model learns to produce utterances word by word: The words of a sentence are passed through a linear word embedding layer and then fed, together with the encoded image features[6], into a one-layer LSTM (Hochreiter and Schmidhuber, 1997).

2.3 Model Training

Perception-based learning is realized by training the model using a cross-entropy loss. The model is given pairs of images with corresponding sentences and uses these to learn a mapping from the visual to the language domain. Given an image i and a target ground-truth sentence s consisting of the words $w_1, \ldots, w_T$, the loss is defined as:

$$\mathcal{L}_{XSL}(\theta) = -\sum_{t=1}^{T} \log p_\theta(w_t | w_{<t}; i) \qquad (1)$$

Production-based learning is instantiated by training the model using REINFORCE (Williams, 1992). To operationalize the Communicative Feedback (i.e., the reward), we calculate the BLEU score (Papineni et al., 2002) between the produced sentence and all 6 reference descriptions/captions from the dataset, taking into account both the quality of semantics as well as word order (n-gram sequences).[7] Crucially, the BLEU score takes into account the fact that there is not only one correct sentence for each image, but rather a range of equally adequate ways to describe the same scene. In particular, if the model produces an exact imitation of one of the reference sentences, it obtains the highest BLEU score, even if the other 5 reference sentences are very different.

Given an image i, the sampled sentence from the model $s_m = w_1, \ldots, w_T$ and the 6 reference sentences $S_{ref} = s_1, \ldots, s_6$, the loss is defined as follows:

$$\mathcal{L}_{CF}(\theta) = -\sum_{t=1}^{T} r(s_m, S_{ref}) \cdot \log p_\theta(w_t) \qquad (2)$$

where $r(s_m, S_{ref}) = BLEU(s_m, S_{ref})$.

More details on model hyperparameters can be found in Appendix B.

2.4 Model Evaluation

In order to evaluate the model's acquisition of visually-grounded semantics, we use an evaluation method proposed by Nikolaus and Fourtassi (2021). It is based on 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). Note that the models

[5] As commonly applied in other multimodal XSL work (Chrupała et al., 2015; Khorrami and Räsänen, 2021).

[6] While Vinyals et al. (2015) fed the image features only at the first timestep into the LSTM, here we feed it at every timestep as this showed to improve performance on our evaluation substantially. An explanation could be that when feeding the image features only at the first timestep the model gradually *forgets* about the input, and relies more on the language modeling task of next-word prediction, which does not aid the learning of *visually-grounded* semantics.

[7] While the BLEU score only measures the adequacy of the children's produced sentence, we used it here as a proxy for adults' Communicative Feedback. The assumption being that the degree to which adults provide positive, contingent responses (i.e., cues of coordination success) depends closely on children's production adequacy as was shown previously, though in a different context, by Warlaumont et al. (2014). We return to this assumption in the Discussion.

	Evaluation task	Accuracy XSL	 Alt	 XSL+CF	 XSL+Alt
Word-level Semantics	Nouns: Persons	**0.87** ± 0.03	0.51 ± 0.01	0.79 ± 0.03	**0.87** ± 0.04
	Nouns: Animals	**0.99** ± 0.01	0.53 ± 0.05	0.98 ± 0.01	**0.99** ± 0.00
	Nouns: Objects	0.94 ± 0.01	0.51 ± 0.01	0.94 ± 0.00	**0.95** ± 0.00
	Verbs	0.55 ± 0.05	0.50 ± 0.00	**0.77** ± 0.04	0.73 ± 0.05
	Adjectives	0.75 ± 0.02	0.50 ± 0.01	0.81 ± 0.03	**0.82** ± 0.02
Sentence-level Semantics	Adj-noun dependencies	0.61 ± 0.03	0.50 ± 0.00	0.62 ± 0.02	**0.63** ± 0.03
	Verb-noun dependencies	0.55 ± 0.03	0.50 ± 0.00	**0.72** ± 0.05	0.68 ± 0.02
	Semantic roles	**0.65** ± 0.07	0.50 ± 0.01	0.61 ± 0.05	0.61 ± 0.07
	Average	0.74 ± 0.01	0.51 ± 0.01	0.78 ± 0.01	**0.79** ± 0.01

Table 1: Accuracy (mean and standard deviation over 5 runs with different random initializations) for all semantic evaluation tasks for different learning scenarios.

are not trained to optimize these tasks. The tasks are only used during the evaluation phase and they test if the models learn various aspects of semantics as a "side product" of XSL and CF. Indeed, when we evaluate children's knowledge in the lab, we do not suppose they have acquired their knowledge by being trained on lab tasks.

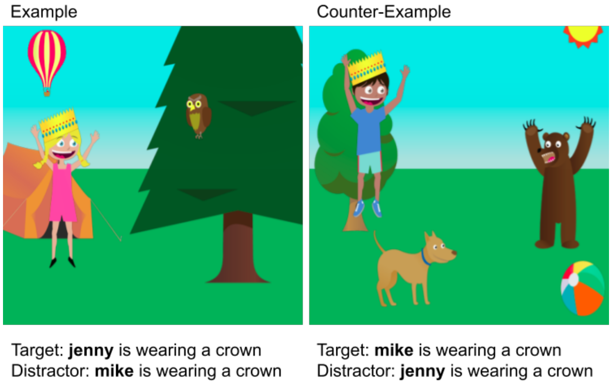

Example
Counter-Example
Target: **jenny** is wearing a crown
Distractor: **mike** is wearing a crown
Target: **mike** is wearing a crown
Distractor: **jenny** is wearing a crown

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. Figure reproduced from Nikolaus and Fourtassi (2021).

These tasks test the model's learning of grounded semantics on the word level (nouns, adjectives, verbs) and sentence level (adjective-noun dependencies, verb-noun dependencies, semantic roles). A task involves multiple test trials, each consists of an image, a target sentence and a distractor sentence: (i, s_t, s_d). Critically, each test trial is *counter-balanced* to control for linguistic biases (e.g., that Jenny occurs most frequently as semantic agent and Mike more as a semantic patient), in a way that a language model that does not have access to the image data performs at chance (see also Figure 1, more examples are shown in Appendix A).[8]

The model's accuracy at choosing the correct sentence s_t given the image i indicates how well it has learned visually grounded semantics for the phenomenon under study. We operationalize the model's choice for a trial by calculating both the perplexity of the target sentence s_t given the image i and the perplexity of the distractor sentence s_d given i. If the perplexity of the target sentence s_t is lower, the trial has been successfully completed.

3 Analyses

3.1 Comparing learning scenarios

We study and compare four different learning scenarios:

XSL: Pure perception-based learning In this scenario, the model learns only using XSL. It represents our baseline against which we compare configurations including CF.

Alt: Alternating between perception and production-based learning Here, the model switches between the XSL and CF objectives throughout the entire learning process.

XSL+CF: First pure perception-based learning, then pure production-based learning We train

[8] Besides controlling for linguistic biases, the evaluation sets also control for some potential visual biases, e.g., that the semantic agent may occur more frequently on the left side of the image (see Nikolaus and Fourtassi (2021) for more details).

the model until convergence using XSL, and afterwards we fine tune the model using CF.

XSL+Alt: First pure perception-based learning, then alternation The model is first trained until convergence using XSL, but afterwards, we alternate between XSL and CF. This scenario is intuitively the most plausible one: Once the language learner starts to speak (i.e. produce their own utterances), this does not mean that they stop to engage in perception-based learning. Rather, they continue learning using both mechanisms.

Accuracies for the four different learning scenarios are reported in Table 1.[9] The scenario XSL learns word-level and sentence-level semantics relatively well compared to the other scenarios. It only appears to struggle with the verbs and the verb-noun dependencies. This fact highlights the role of XSL as a major learning mechanism. When looking at the results of Alt, we can conclude that combining XSL and CF from the start deteriorates the performance (compared to XSL alone) of all metrics. This deterioration was observed regardless of the frequency of alternation between XSL and CF (for direct comparison with XSL+Alt we only report results using one XSL update every 10 CF updates in Table 1, but see Appendix C for results with other alternation frequencies).

Moving to the more plausible scenarios (where production comes into play only after a phase of pure perception-based learning), we found that for XSL+CF, we have, on the one hand, an increase in performance (compared to the baseline XSL) in some categories like "verbs," "adjectives,", and "verb-noun dependencies." On the other hand, we observe a decrease in other categories, especially the category "persons." Finally, the scenario XSL+Alt leads to the best overall results except for verbs and semantic roles, but the difference is within the margin of error. Here we only show results of XSL+Alt using one XSL update every 10 CF updates (which seems to optimize performance), but other – both lower and higher – ratios only marginally change the model's behavior and the conclusions remain the same (see Appendix C).

Appendix D contains a comparison of the BLEU scores (our measure of utterance adequacy) for the different learning scenarios. Consistent with our semantic evaluation results, XSL+Alt leads to the highest BLEU score.

3.2 Developmental Trajectories

Results in Table 1 show evaluation scores after the model has converged on the entire dataset. Here we test the developmental trajectories in each semantic category using different data sizes as a proxy for progression in time. Figure 2 shows the accuracy for different tasks when the best-performing model XSL+Alt is trained on different training data sizes. Already with very small training data (10% of the original training set, 800 examples), nouns and adjectives are learned to a high degree. Verbs and sentence-level semantics are learned only with larger training set sizes.

3.3 Effect of the data size used for XSL pre-training

In the best performing configuration, XSL+Alt, the model was first pre-trained on the entire dataset using XSL, and then trained further using XSL and CF, using again the entire dataset. However, in real life, children spend only a fraction of their learning time (generally the first year of their life) doing pure perception-based learning. Thus, here we test how different fractions of pre-training data influence performance.

Figure 3 shows the average task accuracy (cf. last row in Table 1) for XSL+Alt models that are pre-trained until convergence on training datasets of different size, and then trained in alternation between XSL and CF on the full training dataset until convergence. While the results indicate that more pre-training data is better, we observe a steep gain in average task accuracy starting from pre-training only 5% of the data (up from chance level with 0% pre-training, a limit case that corresponds to the scenario of Alt alone), indicating that even a small amount of perception-based training is useful to initiate a successful learning trajectory.

4 Discussion

How do children learn the meanings of words and sentences in their native language? Previous modeling effort has largely focused on perception-based learning mechanisms such as XSL. However, children do not learn only by mere exposure to the perceptual cross-modal input, they also practice their early – albeit rudimentary– knowledge and receive feedback from caregivers, which allows them to correct/refine this knowledge (Clark, 2018,

[9] Note that the results are not directly comparable to the results for the cross-situational learner in Nikolaus and Fourtassi (2021), see Appendix E for more detail.

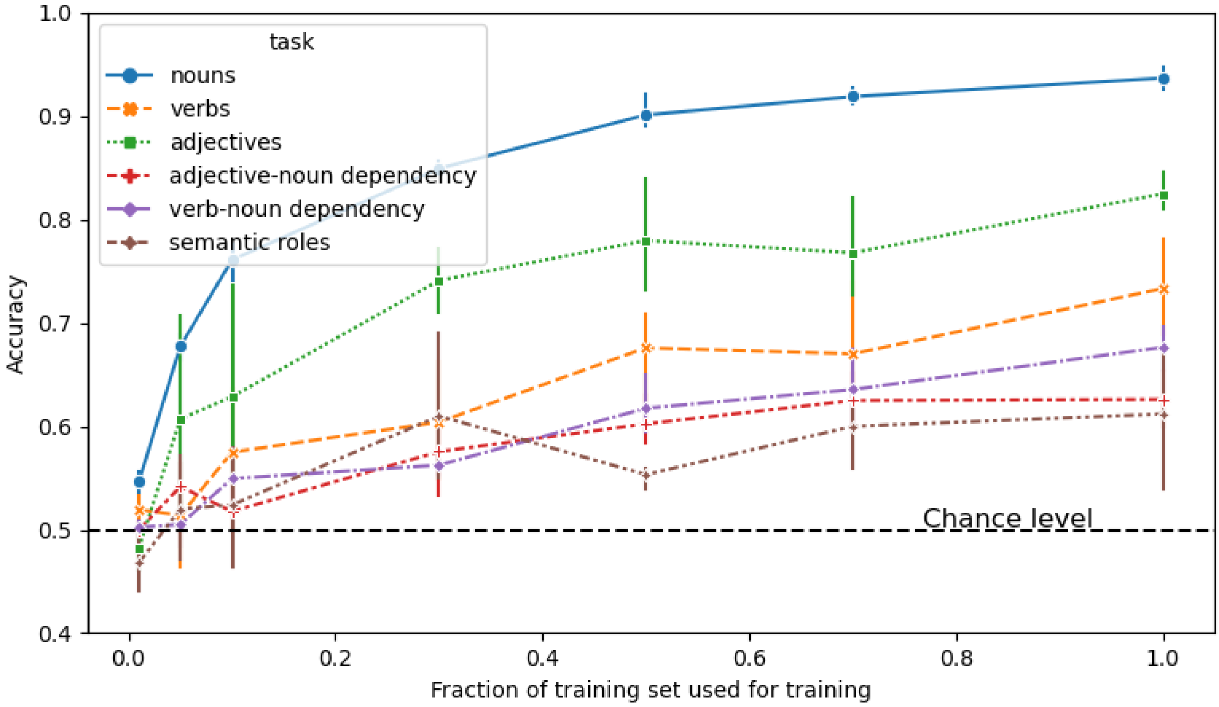


Figure 2: Accuracy as a function of training set size for best performing learning setup (`XSL+Alt`). Vertical bars indicate the standard deviation over 5 runs. Accuracies for all noun categories were averaged.

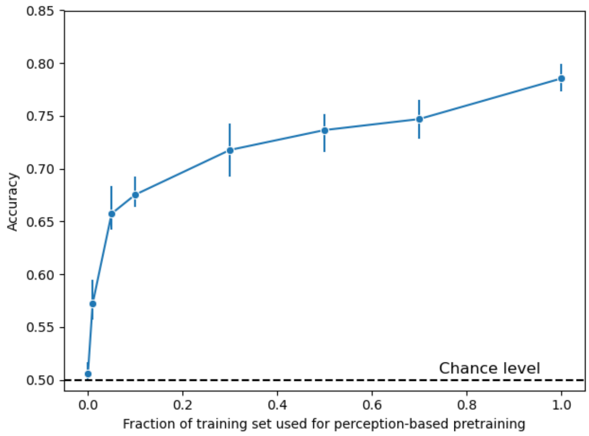


Figure 3: Average accuracy as a function of amount of perception-based pre-training for the best performing learning setup (`XSL+Alt`). Vertical bars indicate the standard deviation over 5 runs.

2020). Here we investigated one possible feedback mechanism on children's early production (CF), that relies on general coordination and miscoordination cues, and does not necessarily require the caregiver providing an explicit correction.

We proposed a computational model that integrates both XSL and CF, allowing us to study how these two mechanisms could interact in early semantic learning. The same model learns both from perceptual input and from feedback on production through reinforcement. We tested various learning scenarios that varied in their plausibility given our understanding of how children's learn language. Crucially, we found that the most plausible learning scenario (i.e., `XSL+Alt`) – where the model first learns through perception, and second through alternating perception and production – is also the one that leads to the best overall performance on most semantic tasks.

The fact that `XSL+Alt` performed better than `XSL` alone confirms the main hypothesis of this work: CF plays a role in semantic learning above and beyond XSL. In addition, the fact that `Alt` – which alternates perception and production from the start – hurts performance compared to XSL, suggests that for CF to be effective, it requires a first phase of learning through perception, which is an intuitive finding since the model has first to be exposed to enough linguistic/semantic input to be able to start producing – at least partially – meaningful utterances (for which RL is more useful). This finding also corresponds to children's learning trajectory where they only start producing words (and receiving feedback on them) after a period of pure perception-based learning.[10]

[10] Children do not generally utter their first words until they are about 10 months old (Frank et al., 2021) while they already understand certain words well before that age (Bergel-

Interactions between perception-based and production-based learning Another interesting finding of this work is that XSL+Alt (e.g., alternating XSL and CF after a period of pure XSL) performs better than XSL+CF (i.e., using CF alone after a period of pure XSL). This finding means that when CF is combined with XSL, it leads to improvement in performance compared to when either XSL or CF operates alone or in a sequential fashion. In other words, we found that XSL and CF interact *synergistically* to improve performance. In what follows, we examine this observed synergy in more details.

Results in Table 3 show that while XSL+CF improved performance on "verbs" compared to XSL, it also led to a significant drop in the category "persons."[11] We speculate that by using reinforcement learning alone, XSL+CF explores the hypothesis space and picks short utterances that lead to a high reward signal and continues (re)producing them. While this behavior could lead to improvement for the parts of the language that are well covered by this local space (e.g., verbs), it can also lead to a drop in performance for the other aspects. In particular, here the difference between Jenny and Mike in the category "persons" may become forgotten.

Qualitative and quantitative investigation of the model's behavior supports our speculation. For example, when we sample sentences randomly from the productions of XSL+CF and XSL+Alt given images in the validation set, we observed that while XSL+CF produces a variety of verbs (similar to XSL+Alt), it tends to produce systematically shorter utterances involving disproportionately only one person (see Table 6 in Appendix F).

Figure 4 confirms this observation quantitatively: XSL+CF increasingly produces sentences involving Jenny, but decreasingly sentences involving Mike. This fact leads to the situation where the model gets less feedback on the difference between Jenny and Mike and, therefore, *unlearns* this distinction to some degree.

For XSL+Alt, the fraction of sentences involving Jenny and Mike remains largely constant, thus avoiding the problem faced by XSL+CF. At the same time, XSL+Alt keeps a balanced coverage of verbs allowing it to maintain the good scores achieved by XSL+CF on this category (see Appendix F for a quantitative analysis comparing the production of verbs in both models).

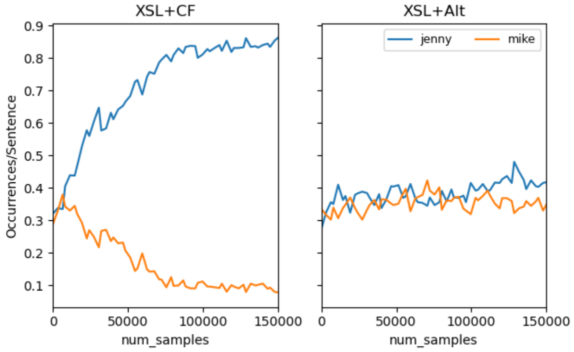


Figure 4: Comparison of the fraction of occurrences of persons ("jenny" and "mike") in sentences produced during training of the XSL+CF (left) and XSL+Alt (right) training setups. The graphs only display the second training step, not the pre-training using XSL.

The conclusion we draw from comparing XSL+CF and XSL+Alt is that, even after a period of pure XSL, continuing to learn through XSL from time to time while doing reinforcement on production helps the model not to get biased towards a subset of the language it is supposed to learn. Similar phenomena of "language drift" – due to reinforcement learning operating alone – have been observed in another line of work studying emergent communication systems (Lewis et al., 2017; Lowe et al., 2019; Lazaridou et al., 2020).

Learning Trajectories The best performing model, i.e. XSL+Alt, not only instantiates – intuitively – the most plausible learning scenario in early childhood, it also recapitulates some specific findings in the language development literature about the timeline of semantic learning. For example, it learns nouns before predicates (adjectives and verbs), resonating with previous findings about the "noun bias" (Gentner, 1982; Bates et al., 1994; Frank et al., 2021). That said, the models' performance on verbs (relative to other parts of speech) should be interpreted with caution given the fact that we only used static images in both training and testing. In real life, children learn verbs from dynamic actions and some experimental studies also evaluate verb learning use videos instead of static images (Golinkoff et al., 1987; Gertner et al., 2006).

son and Swingley, 2012), indicating that they engage in a perception-based learning well before starting to produce their own utterances.

[11] The drop in "persons" could explain the slight drop in "semantic roles," (as distinguishing the persons is a prerequisite to understand semantic roles) however this slight drop is within the margin of error, so we could not draw strong conclusions about the difference with XSL for this category.

The model shows a rather late onset of understanding sentence-level semantics such as semantic roles, only after a sizable lexicon has been acquired. This fact mirrors, e.g., the finding that children show evidence of recognize semantic roles in a sentence during their second year of life (Golinkoff et al., 2013), that is, at an age when they have already acquired a substantial vocabulary (Frank et al., 2021). Note that the model's performance on sentence-level semantics remains relatively low compared to word-level semantics even when learning from the entire dataset. It is difficult, based only on the current results, to conclude whether more data will lead to improvement in sentence-level semantics or whether the model has already reached its ceiling performance due to structural limitations (e.g., the lack of higher-level conceptual knowledge about semantic agency).

Limitations and future research directions While our modeling work has allowed us to test crucial hypotheses about semantic learning, it used – like any modeling work – simplifying assumptions about the phenomenon under study. For example, here we used an integrated model for both perception and production. This choice was primarily motivated by parsimony. While it allowed us to provide a direct comparison of XSL and CF, it abstracted away limitations in children's production abilities compared to perception (e.g., due to immature motor/articulatory skills) and from difficulties that children face when trying to coordinate production with perception (e.g. Clark and Hecht, 1983). In addition, we did not account for constraints on children's information processing abilities during the learning process (e.g., limited attention span and working memory), and how these constraints may, for example, translate in the learner focusing on specific parts of the input (Gelderloos et al., 2020).

More generally, the current work focused on investigating the input-output mapping problem for semantic learning and how Communicative Feedback can help such learning. It did not intend to account for the exact cognitive processes that operate in children's mind nor did it take into account specific cognitive limitations and constraints when trying to achieve this mapping. Thus, this work is best situated at the computational level of analysis (Marr, 1982), which is a necessary first step towards a deeper understanding of the cognitive implementation.

Another simplifying assumption of this work was the use of the BLEU score as a reward to the model when learning through reinforcement. In other words, we used a measure that only evaluates the extent to which the learner's utterance is correct as a proxy for how the teacher would react. While this assumption is grounded in previous experimental work showing that adults' responses are contingent on children's type of vocalization (Warlaumont et al., 2014), here we went beyond the broad distinction studied in this previous work (speech vs. non-speech) and assumed that adults' responses are also contingent on the adequacy of speech itself. That is, immediate, positive reaction from adults is more likely to follow correct/adequate speech from the child, which would encourage the re-use of adequate (but not inadequate) speech in subsequent conversations.

Note that the BLEU score feeds the model with ideal information whereas the feedback that children receive in real life is highly dynamic, multimodal and noisy. While, as we said above, the current paper took a computational level of analysis approach that only studied learning under optimal conditions, future work is required to (1) estimate the quality and frequency of Communicative Feedback in child-caregiver conversations (CHILDES (MacWhinney, 2000)) and (2) use these findings to assess the scalability of the current proposal to account for child's language use and development in the real world.

In conclusion, this paper provides a quantitative proof of concept about the role production-based learning can play in semantic knowledge acquisition together with perception-based learning. An important finding was that combining both mechanisms leads to synergistic learning. One question for future experimental work is whether such synergy can be observed in controlled behavioral experiments.

Acknowledgements

We thank Mostafa Abdou 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).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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.

Kathleen Bloom. 1988. Quality of adult vocalizations affects the quality of infant vocalizations. *Journal of Child Language*, 15(3):469–480.

Roger Brown. 1958. *Words and things.* Free Press.

Roger Brown. 1973. *A first language: The early stages.* Harvard U. Press.

Marisa Casillas and Elma Hilbrink. 2020. 3. communicative act development. In *Developmental and Clinical Pragmatics*, pages 61–88. De Gruyter Mouton.

Jane B Childers, Julie Vaughan, and Donald A Burquest. 2007. Joint attention and word learning in ngas-speaking toddlers in nigeria. *Journal of Child Language*, 34(2):199.

Michelle M Chouinard and Eve V Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3):637–670.

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.

Eve V Clark. 2018. Conversation and language acquisition: A pragmatic approach. *Language Learning and Development*, 14(3):170–185.

Eve V Clark. 2020. Conversational repair and the acquisition of language. *Discourse Processes*, 57(5-6):441–459.

Eve V Clark and Barbara F Hecht. 1983. Comprehension, production, and language acquisition. *Annual review of psychology*, 34(1):325–349.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Cynthia Fisher and Lila R Gleitman. 2002. *Language acquisition.* John Wiley & Sons Inc.

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

Lieke Gelderloos, Alireza Mahmoudi Kamelabad, and Afra Alishahi. 2020. Active word learning through self-supervision. In *CogSci*.

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.

Yael Gertner, Cynthia Fisher, and Julie Eisengart. 2006. Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological science*, 17(8):684–691.

Michael H Goldstein, Andrew P King, and Meredith J West. 2003. Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences*, 100(13):8030–8035.

Michael H Goldstein and Jennifer A Schwade. 2008. Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological science*, 19(5):515–523.

Roberta Michnick Golinkoff, Kathryn Hirsh-Pasek, Kathleen M Cauley, and Laura Gordon. 1987. The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of child language*, 14(1):23–45.

Roberta Michnick Golinkoff, Weiyi Ma, Lulu Song, and Kathy Hirsh-Pasek. 2013. Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, 8(3):316–339.

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.

Sarah Hiller and Raquel Fernández. 2016. A data-driven investigation of corrective feedback on subject omission errors in first language acquisition. In

Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 105–114, Berlin, Germany. Association for Computational Linguistics.

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.

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

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. 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.

Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. 2019. On the interaction between supervision and self-play in emergent communication. In *International Conference on Learning Representations*.

Brian MacWhinney. 2000. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.

Gabriela Markova and Maria Legerstee. 2006. Contingency, imitation, and affect sharing: Foundations of infants' social awareness. *Developmental psychology*, 42(1):132.

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

Judi Mesman, Tessa Minter, Andrei Angnged, Ibrahima AH Cissé, Gul Deniz Salali, and Andrea Bamberg Migliano. 2018. Universality without uniformity: A culturally inclusive approach to sensitive responsiveness in infant caregiving. *Child Development*, 89(3):837–850.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China. Association for Computational Linguistics.

Mitja Nikolaus and Abdellah Fourtassi. 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.

Mitja Nikolaus, Juliette Maes, Jeremy Auguste, Laurent Prevot, and Abdellah Fourtassi. 2021. Large-scale study of speech acts' development using automatic labelling. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

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.

D Kimbrough Oller. 2000. *The emergence of the speech capacity*. Psychology Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

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.

Matthew Saxton, Carmel Houston-Price, and Natasha Dawson. 2005. The prompt hypothesis: Clarification requests as corrective input for grammatical errors. *Applied Psycholinguistics*, 26(3):393–414.

Linda Smith and Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Catherine E Snow, Barbara Alexander Pan, Alison Imbens-Bailey, and Jane Herman. 1996. Learning how to say what one means: A longitudinal study of children's speech act use. *Social Development*, 5(1):56–84.

Edward Tronick, Heidelise Als, Lauren Adamson, Susan Wise, and T Berry Brazelton. 1978. The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child psychiatry*, 17(1):1–13.

Sho Tsuji, Alejandrina Cristia, and Emmanuel Dupoux. 2021. Scala: A blueprint for computational models of language acquisition in social context. *Cognition*, page 104779.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

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*.

Wai Keen Vong and Brenden M Lake. 2021. Cross-situational word learning with multimodal neural networks.

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*.

Anne S Warlaumont, Jeffrey A Richards, Jill Gilkerson, and D Kimbrough Oller. 2014. A social feedback loop for speech development and its reduction in autism. *Psychological science*, 25(7):1314–1324.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

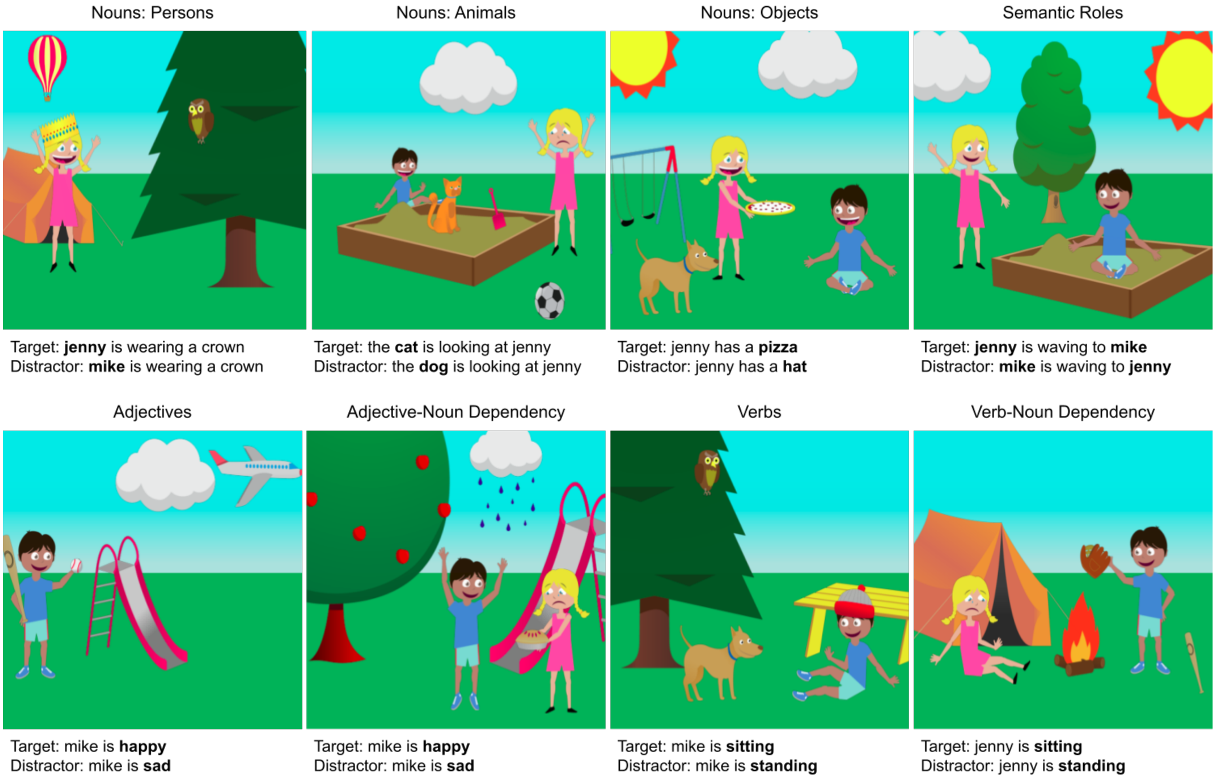
Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

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.

A Semantic evaluation examples

For reference, Figure 5 shows an example for each different semantic evaluation set as proposed in Nikolaus and Fourtassi (2021).



Nouns: Persons

Target: **jenny** is wearing a crown
Distractor: **mike** is wearing a crown

Nouns: Animals

Target: the **cat** is looking at jenny
Distractor: the **dog** is looking at jenny

Nouns: Objects

Target: jenny has a **pizza**
Distractor: jenny has a **hat**

Semantic Roles

Target: **jenny** is waving to **mike**
Distractor: **mike** is waving to **jenny**

Adjectives

Target: mike is **happy**
Distractor: mike is **sad**

Adjective-Noun Dependency

Target: mike is **happy**
Distractor: mike is **sad**

Verbs

Target: mike is **sitting**
Distractor: mike is **standing**

Verb-Noun Dependency

Target: jenny is **sitting**
Distractor: jenny is **standing**

Figure 5: Examples for the evaluation of word and sentence-level semantics. Each test trial consists of an image, a target and a distractor sentence. Reproduced from Nikolaus and Fourtassi (2021).

B Hyperparameters

Model hyperparameters as indicated in Table 2 were chosen based on general best-practices and not any further tuned (except for the frequency of CF updates, see Appendix C). During training, we evaluate the model every 100 batches, and stop training if the BLEU score on the held out validation set does not improve for 50 consecutive validations. All models converged within 8 hours when running on a single GPU.

Parameter	**Value**
Minimum word frequency for inclusion vocab	5
Word Embeddings Size	100
LSTM Hidden Layer Size	512
Optimizer	Adam
Optimizer Initial Learning Rate	$1 \cdot 10^{-4}$
Optimizer Initial Learning Rate (Model fine-tuning)	$1 \cdot 10^{-5}$
Dropout	0.2
Batch size	32

Table 2: Model hyperparameters.

C Varying frequency of CF updates

As the loss terms of the cross-entropy loss used in XSL and the policy gradient loss used in CF can take very different margins, we experiment with different update frequencies of XSL updates with respect one XSL update. An update frequency of 2 indicates that we perform an XSL update every 2 CF updates.

The results as shown in Table 3 show that we obtain the best results (average over all tasks) when performing 1 XSL update every CF update for the model in the `Alt` setup, that is when alternating production-based and perception-based learning from the start. However, the performance is still worse than for a model trained using `XSL` alone (mainly regarding persons and semantic roles).

For our best performing setup, `XSL+Alt`, we observe a different pattern, displayed in Table 4. In this case it is best to perform an XSL update every 10 CF updates. We hypothesize that this can be explained by the fact that the CF updates are more useful in this setup, as the model has already learned a language model in the first perception-based learning phase before starting to produce sentences. In the main text, we report results for both `Alt` and `XSL+Alt` with a frequency of 10 CF updates per XSL update for direct comparison.

		Frequency of CF updates				
Evaluation task		1	2	5	10	20
Word-level Semantics	Nouns: Persons	0.740	0.660	0.520	0.500	0.480
	Nouns: Animals	0.997	0.978	0.703	0.667	0.500
	Nouns: Objects	0.930	0.858	0.720	0.567	0.497
	Verbs	0.597	0.556	0.542	0.486	0.500
	Adjectives	0.786	0.714	0.643	0.554	0.500
Sentence-level Semantics	Adj-noun dependencies	0.786	0.714	0.643	0.554	0.500
	Verb-noun dependencies	0.565	0.573	0.542	0.510	0.500
	Semantic roles	0.540	0.500	0.480	0.500	0.440
	Average	**0.715**	0.674	0.588	0.537	0.490

Table 3: Accuracy for all semantic evaluation tasks for varying frequency of CF updates in the `Alt` setup. Note that we only performed one run for each setting, and thus some numbers do not match exactly those in the Table 1.

		Frequency of CF updates				
Evaluation task		1	2	5	10	20
Word-level Semantics	Nouns: Persons	0.900	0.880	0.880	0.860	0.900
	Nouns: Animals	0.997	0.994	0.997	0.997	0.997
	Nouns: Objects	0.952	0.957	0.954	0.954	0.949
	Verbs	0.722	0.708	0.764	0.778	0.764
	Adjectives	0.750	0.857	0.786	0.839	0.839
Sentence-level Semantics	Adj-noun dependencies	0.646	0.667	0.630	0.594	0.635
	Verb-noun dependencies	0.598	0.593	0.630	0.720	0.708
	Semantic roles	0.620	0.620	0.680	0.660	0.480
	Average	0.773	0.785	0.790	**0.800**	0.784

Table 4: Accuracy for all semantic evaluation tasks for varying frequency of CF updates in the `XSL+Alt` setup. Note that we only performed one run for each setting, and thus some numbers do not match exactly those in the Table 1.

D BLEU Scores

Table 5 shows the BLEU scores for all different learning scenarios. The score was calculated by sampling images from the validation set and comparing generated sentences with the gold sentences. These results are compatible with our observations using the grounded semantics evaluation tasks. Here again `XSL+Alt` performs best.

`XSL`	`Alt`	`XSL+CF`	`XSL+Alt`
66.5 ± 0.8	53.9 ± 0.6	70.8 ± 0.2	$\mathbf{72.7} \pm 0.5$

Table 5: BLEU score on the test set (mean and standard deviation over 5 runs) for different learning setups.

E Comparison with Nikolaus and Fourtassi (2021)

Our baseline (`XSL`) results differ from the results in Nikolaus and Fourtassi (2021) for several reasons.

Firstly, their models are trained with a max-margin loss, instead of a cross-entropy objective as we did here. We cannot evaluate our model by directly calculating similarity between images and sentences because it does not learn a multimodal semantic embedding space. Thus, we evaluate it by calculating conditional perplexity for both target and distractor sentences. These factors might explain the drop in performance for some metrics, especially for sentence-level semantics. Future work should investigate how to combine both training objectives (max-margin loss and cross-entropy loss), in order to combine their respective benefits (e.g. Nikolaus et al., 2019).

Secondly, we do fine-tune the ResNet of our models, as we observed substantial performance improvements with this change. This might explain the gain in performance for adjectives (the children's emotions), which the model of Nikolaus and Fourtassi (2021) struggled with (probably due to the inappropriateness of the pre-trained image features, they are largely optimized for recognizing objects in naturalistic scenes, but not clip-art objects).

F Analysis of produced sentences

Examples of models' produced sentences (at the end of training) are shown in Table 6.

`XSL+CF`	`XSL+Alt`
jenny is wearing glasses	jenny is crying
an owl is sitting	mike is holding balloons
jenny is holding	mike is kicking the soccer ball
jenny is holding balloons	jenny is holding a ketchup
jenny is flying	jenny is playing in the sandbox
jenny is holding the	jenny has glasses on
jenny is holding	mike is making a pirate
jenny is wearing	jenny is running away from the snake
mike is wearing	the bear is wearing a wizards hat
jenny is angrily	the rain is cooking lightning in the sky

Table 6: 10 sentences produced by the models for randomly sampled images from the validation set. The model checkpoints used were from the end of training (epoch 19).

We further quantitatively compare the produced utterances during the training using `XSL+CF` and `XSL+Alt`. Every 100 batches, we sample sentences from the model for all images in the validation set and analyze these produced sentences for sentence length (Figure 6) as well as occurrences of persons (Figure 4) and verbs (Figure 7). There are only 2 persons in the dataset, "jenny" and "mike". We measure occurrence of persons by counting sentences that contain "jenny", but not "mike" (and vice versa). Regarding the verbs, we count occurrences for all verbs that are used in the semantic evaluation tasks.

The examples show that the model produces increasingly short sentences when trained using `XSL+CF`. We also observe a drop in mean sentence length for `XSL+Alt`, but to a substantially smaller degree.

Figure 4 shows that the model trained using XSL+CF increasingly produces sentences involving "jenny", but decreasingly sentences involving "mike". Thus it might get less feedback on the difference between Jenny and Mike and unlearn this distinction to some degree. Consequently, it also struggles more to understand semantics roles (distinguishing the persons is necessary to correctly map the semantic roles). For XSL+Alt, the fraction of sentences involving "jenny" and "mike" remains largely constant.

Regarding the presence of verbs, Figure 7 shows a different pattern. While for XSL+Alt the fractions do not vary much, in XSL+CF some verbs are produced increasingly. This might explain the large gain in performance for verbs: The model produces more sentences involving verbs, and thus also receives more valuable feedback to learn meaningful semantic representations.

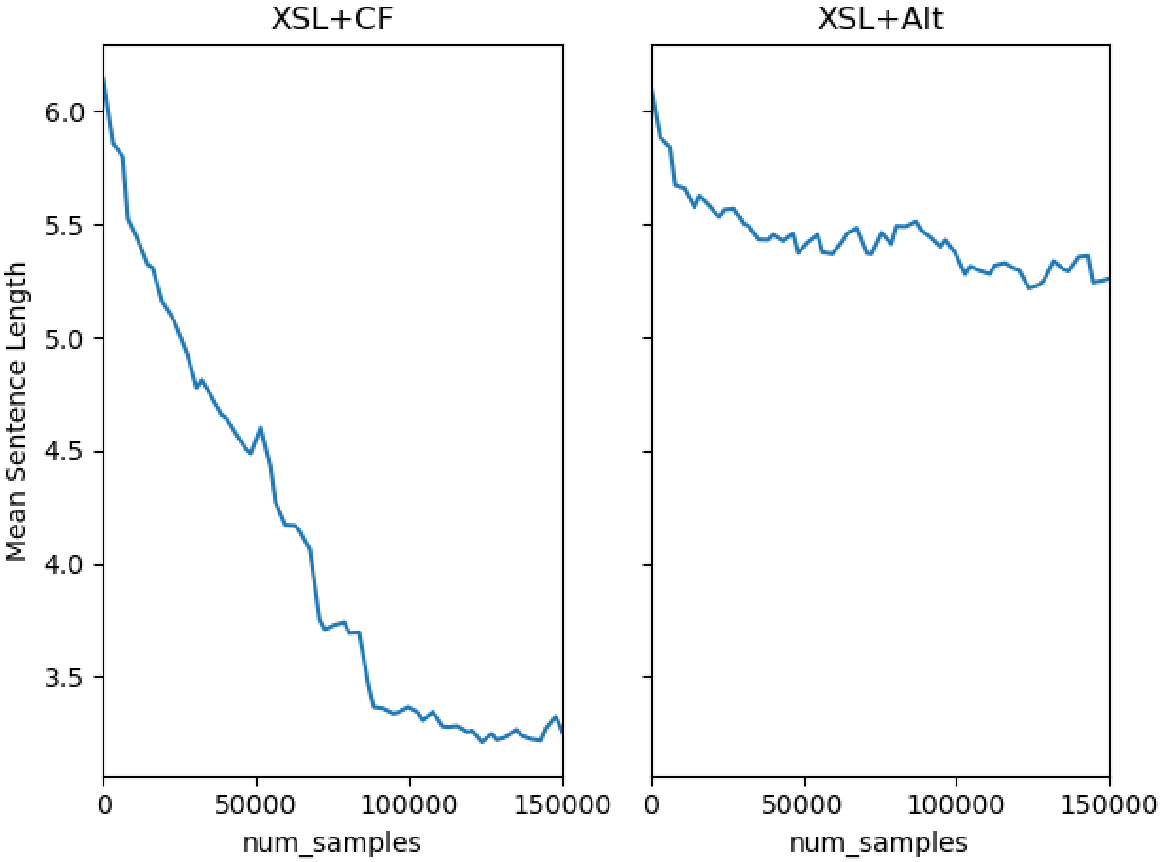


Figure 6: Comparison of the mean sentence length during training of the XSL+CF and XSL+Alt training setups. The graphs only display the second training step, not the pre-training using XSL.

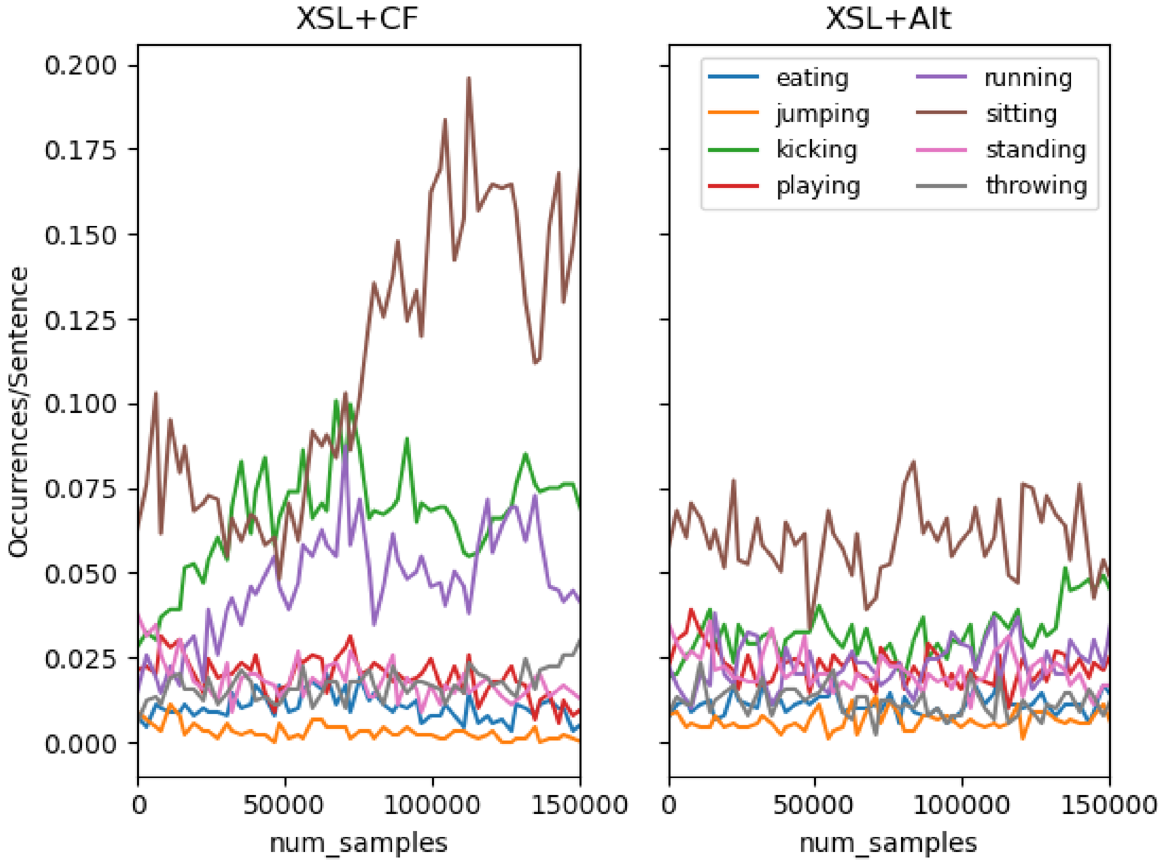


Figure 7: Comparison of the fraction of occurrences of verbs during training of the `XSL+CF` and `XSL+Alt` training setups. The graphs only display the second training step, not the pre-training using XSL.