Curious Topics: A Curiosity-Based Model of First Language Word Learning

Daan Keijser (daankeijser@icloud.com)
Department of Cognitive Science and Artificial Intelligence, Tilburg University
Warandelaan 2, 5037 AB Tilburg

Lieke Gelderloos (l.j.gelderloos@uvt.nl)
Department of Cognitive Science and Artificial Intelligence, Tilburg University
Warandelaan 2, 5037 AB Tilburg

Afra Alishahi (a.alishahi@uvt.nl)
Department of Cognitive Science and Artificial Intelligence, Tilburg University
Warandelaan 2, 5037 AB Tilburg

Abstract

This paper investigates whether a curiosity-based strategy could be beneficial to word learning. Children are active conversation partners and exert considerable influence over the topics that are discussed in conversation with their parents. As the choice of topics is likely to be intrinsically motivated, a formalization of curiosity is implemented in a word learning model. The model receives annotated Flickr30k Entities images as input, and is trained in two conditions. In the curious condition, the model chooses objects to talk about from the scene according to the curiosity mechanism, whereas in the random condition, the model receives randomly chosen objects as input. The goal of this study is to show how a curious, active choice of topics by a language learner improves word learning compared to random selection. Curiosity is found to make word learning faster, increase robustness, and lead to better accuracy.

Keywords: word learning; curiosity; interaction; connectionist model.

Introduction

Language learning research focuses more and more on child-parent interaction and the social aspects of early conversation. Children are active learners and have considerable agency as conversational partners. We will argue that curiosity is a plausible mechanism for the child to come up with new topics to talk about within this conversational context. While AI researchers have become inspired by the curiosity displayed by children, and have implemented intrinsically motivated exploration in computer models, this formalized curiosity has not been applied to computational models of language learning. At the same time, the implementations of curiosity in computer models are often not cognitively plausible or the degree of plausibility is unknown (as in reinforcement learning), or the input to the model lacks the complexity of the stimuli encountered by the word learner.

Curiosity can be seen as a viable mechanism in language learning if it provides an advantage to the word learning child. In order to see whether curiosity is beneficial to the word learning process, we propose a curiosity-based model that chooses which object in a scene to talk about next. The model chooses its object of interest from among a number of objects in an image, and triggers the adult to provide linguistic input related to that object. The curiosity mechanism suggested by Twomey and Westermann (2018), which maximizes the product of subjective novelty and plasticity, was implemented to select the objects. To reflect the complexity of visual scenes encountered by the child, the model takes Flickr30k images as input, which depict everyday scenes and objects and have been annotated with captions. The accuracy and loss of the model with a curiosity-based selection of topics were compared to those of a model that received the topics randomly.

Related Work

Interaction and Intentionality

Given the social nature of early conversation, language should not be seen as a product but as a dynamic system for communication (Clark, 2016). Language is used and learned in order to convey and receive information. This means that the child is a conversation partner first, and a language learner second. Furthermore, young children are active speakers and language learners. Bloom et al. (1996) observed that children aged 9 through 24 months are most likely to speak first in conversation with their mother, and the mother to speak after the child. Their evidence did not support the scaffolding model, in which the parent takes a prominent role in the conversation by providing a framework that controls the elements beyond the capacity of the learner and lets the learner concentrate on those elements they are capable of producing. Rather, children initiate conversations and, as shown in several studies (Chapman, Miller, MacKenzie & Bedrosian, 1981; Bloom et al., 1996), mothers are likely to adopt the topic proposed by the child, and continue to talk about it.

These studies show a pattern of turn-taking with a clear role division. Often, the child wants to discuss a certain topic and starts by talking about it. The parent makes sure they understand what the child is referring to by rephrasing what the child has said, which functions as feedback to the

language-learning child at the same time (Chouinard and Clark, 2003). The child then assesses whether the parent has understood the initial message, after which the conversation can continue. When children initiate conversations and their parents adopt the proposed topics, children can exert considerable influence on the topics that are discussed and consequently on the feedback they receive.

Because children initiate conversations, and continue discussing the topic when they feel they have been understood, their choice of topics is unlikely to be random. As the language learner decides on the topic themself, taking in the current surroundings and situation, the choice is likely to be intrinsically motivated. Our study investigates whether a curiosity-based selection of the topics to be discussed enhances word learning through comprehension of the symbol-referent pair.

Curiosity

Curiosity is a form of intrinsic motivation. Intrinsic motivation can be defined as doing "an activity for its inherent satisfaction rather than for some separable consequence. When intrinsically motivated, a person is moved to act for the fun or challenge entailed rather than because of external products, pressures, or rewards" (Ryan and Deci, 2000, p. 56). In the 1950s, psychological research assumed that human behavior is mostly extrinsically motivated, by physical drives such as those to alleviate hunger and minimize pain. A major shortcoming of this theory was that it did not account for exploratory and other curious behavior in humans and animals—behavior that does not seek immediate reward (Oudeyer & Kaplan, 2007).

When formalized to be programmed into a computer model or robot, intrinsic motivation and curiosity are often conflated (e.g. Pathak et al., 2017). Intrinsic motivation has mostly been applied in reinforcement learning, providing agents (robots and models) with an intrinsic desire to explore their environments and build better models and representations of them (Schmidhuber, 2010). Studies that implemented intrinsic motivation have shown that intrinsically motivated exploration increases the performance of a model when generalizing to other tasks (Pathak et al., 2017), and this is likely to be the case for humans as well (Twomey & Westermann, 2018).

Reinforcement learning implements a variety of formalizations of intrinsic motivation, such as maximizing the decrease of prediction errors, maximizing or minimizing predictability, or choosing the action that maximizes the agent's ability to perform a task. Some approaches use predefined rewards or external signals that provide feedback on motor functions, both of which are certainly not cognitively plausible. Of other approaches, it is simply not known how cognitively plausible they are (Oudeyer & Kaplan, 2007; Twomey & Westermann, 2018). In fact, not a lot is known for certain about the workings of curiosity in human cognition in general and children's cognitive development in particular.

What is clear is that children are natural explorers, displaying a novelty preference from an early age. Novel stimuli have most potential to yield new insights upon exploration, as little is known about them yet. As a stimulus is perceived, it becomes less interesting over time (habituation), and other stimuli become more interesting relative to the current stimulus as they remain novel when not examined (Mather, 2013).

Under various circumstances, however, children display familiarity preferences. While completely novel stimuli leave a lot to be explored, they can be uninteresting nonetheless as they differ greatly from the child's state of knowledge. Some have suggested that a moderate discrepancy between a stimulus and the child's representation of it could define the optimally interesting stimulus. What moderate means in this context, however, is not a trivial question. How familiarity and novelty preferences influence learning is little understood as of yet (Mather, 2013).

In a recent publication on curiosity-based categorization in infants, Twomey and Westermann (2018; henceforth T&W) simulate infant categorization using an autoencoder, a model that learns to reproduce the input after reducing it to a compact representation. They defined curiosity as maximizing

$$(i - o)o(1 - o) \qquad (1)$$

where i stands for the model input and o for the model output. $(i - o)$ reflects the difference between the input and the output, which is the error of the autoencoder in response to a particular stimulus. $o(1 - o)$ is the derivative of the sigmoid activation function. As such, this part of the formula reflects the potential update made to the model in response to this stimulus, when it is trained using gradient descent. The formula favors stimuli which the model is predicting least accurately (the difference between input and output is large), and stimuli where a small adjustment in representation has the greatest effect on the prediction in terms of accuracy (the sigmoid derivative is large). In T&W, the curiosity condition learned the most robust category, followed by the objective complexity condition.

T&W provide a cognitively plausible mechanism of curiosity, that produced results that fit their empirical data well. That the implementation of curiosity outperformed the other three mechanisms shows that a learner would benefit from applying this strategy. The inputs used in the study are very interpretable, but also rather simple, consisting of eight training instances and three test instances that differed on four features. The present model will use the same curiosity mechanism, and see how it performs when provided with more complex input, consisting of a sizable set of images to approximate the complexity of the language learner's surroundings.

The model of T&W went through the stimuli without replacement, so that the model encountered every stimulus once per epoch. A drawback of this setup is that it does not correspond to how children encounter stimuli in real life, as children have no control over the order in which stimuli are

presented to them. It is also unlikely for children to come across a string of examples of a certain category presented one after the other. Objects and living things are often seen in isolation from other category members, and amid objects of a wide variety of other categories. Our model was therefore presented scenes containing multiple objects it could choose from. The model would pick one object, skipping the other objects as it went on to the next scene. It was free to look at the same or any other object in the scene during the next epoch, meaning that some objects could be ignored altogether. This made the input sequences of our two experimental conditions more different, and perhaps less comparable than in T&W's case, but it also better approximated a word learning context, in which only certain aspects of a scene are in focus at any time.

Methodology

Model

Our language learner model is inspired by a model of referential expression resolution (Rohrbach et al., 2017), which incorporates an expression generation module as well as the main expression resolution component, which allows it to learn under self-supervision. We implement a similar complementary setup, consisting of a listener and a speaker module. The listener represents a child learning which words represent which objects in the visual modality, by receiving linguistic input from an oracle, which represents an adult conversation partner. The listener learns through supervision, comparing the true referent of a word to the referent it expected, and updating its language knowledge accordingly.

The incorporation of a speaker module in principle allows the model to be used in a conversational set-up, but in the current work, the emphasis is on comprehension. As we describe in more detail in the section on 'Curiosity', the model's curiosity about an object is calculated based on the ability of the listener to comprehend the label the speaker would give it. The oracle labels the object the learner model is most curious about. In analogy, a parent might name an object their child points out. Learning, however, is not simply mapping the label to the correct object: just like in the random condition, the model learns by predicting the referent of the given word and getting feedback on this prediction. The curiosity mechanism affects only the order the stimuli are presented in, but not the learning process itself. Figure 1 illustrates the architecture of the model. The listener learns to map a given word to its referent in the visual context. A visual scene consists of a number of objects. We extract a visual feature vector for each object using the VGG-16 object recognition model presented by Simonyan & Zisserman (2015), pretrained on ImageNet. We use the last fully connected 4096-dimensional layer, which contains high-level visual information. For each object in a given scene, the embedding of the word given by the oracle was concatenated to the object representation, which was input to the listener. The listener further consists of a 256-unit hidden layer followed by a sigmoid activation function, which is fully connected to a single output unit, also followed by sigmoid activation. Softmax applied to the concatenation of the output values for all the objects in a scene gives a distribution reflecting the probabilities of each object being the referent. The listener was trained under supervision using cross-entropy loss on the concatenated output values. The loss function is a quantification of how far off the model's prediction is from the actual target distribution. Hence, a lower loss value means a better performing model.

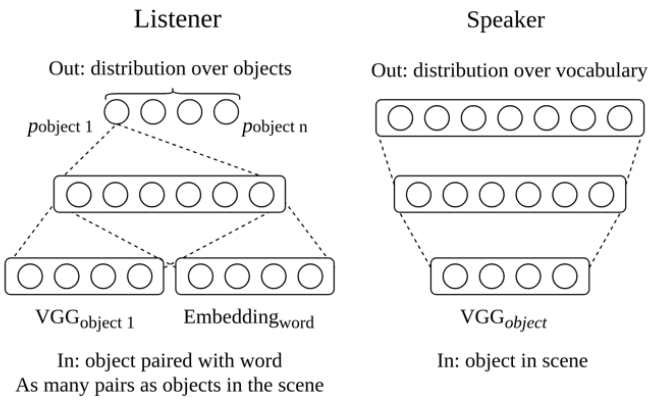

Figure 1: Simplified graphical representation of the model.

The speaker module learns to output a word, given an object. Input to the speaker is a VGG vector, which is fed to a 256 unit hidden layer followed by sigmoid activation, and fully connected to the vocabulary-sized output layer. The speaker was trained using cross-entropy in a self-supervised manner. Rather than training on a single object VGG vectors, it was fed the sum of the VGG vectors of all objects in the scene, weighted by the Softmaxed output vector of the listener (using it as attention). The self-supervision signal consists of the original input word to the listener. Therefore, the speaker can be thought of as learning in an unsupervised manner, although its performance is dependent on that of the listener, which is trained under supervision.

The model was trained using Adam optimization (Kingma & Ba, 2014) in batches of 40 images, for a maximum of 40 epochs. To decide on an initial learning rate, we ran both the 'curious' and the 'random' model, with learning rates ranging from .1 to .00001 for 20 epochs. We ran each condition-learning rate combination with 5 different random initializations. We found that the best scores on the validation data were sometimes obtained in epoch 20, which suggested the model might not have fully converged yet. We therefore decided to report on models trained for 40 epochs. A learning rate of .001 yielded the best results on validation data for both the listener and the speaker. The results reported reflect 20 different runs of both conditions, with learning rate set to .001. The model was implemented in PyTorch (Paszke et al., 2017). The code is available at https://github.com/DaanKeijser/Curious-Topics.

Original captions

A little boy is looking out the balcony surrounded by plants, a toy bike, and plant pots.

A very young boy is looking over the balcony by standing on one of his toy bikes.

A child views the world from their upstairs balcony.

A little boy standing on a plant decorated balcony.

A young boy looks over a white metal balcony.

Selected words

Boy
Balcony
Toy

Figure 2: Example image with captions and selected words.

Data

The Flickr30k dataset (Young et al., 2014) was used as visual input to the model. The dataset consists of 31,783 images taken from Flickr, annotated with five captions per image (158,915 in total) via crowdsourcing. The images depict everyday activities and scenes. Plummer et al. (2015) expanded the dataset with Flickr30k Entities, by identifying which words in the captions refer to which entities in the images. They provided annotation for 244,035 such coreference chains, and located the entities they referred to in the images, resulting in 275,775 bounding boxes. It should be noted that this data has a high level of complexity, but the captions are not child-directed speech.

Figure 2 gives an example of the data our model was trained on. On the visual side, we simplified the learning problem by excluding any referring expressions that described multiple objects, such as 'plants' and 'pots' in Figure 2. Processing multi-word expressions requires a recurrent neural network and a cross-situational learning model, which is outside the scope of the current work. We therefore simplified the referring expressions to single words. The Flickr30k Entities "Sentences" files containing the annotated captions for each image were searched to find all descriptions for every object ID. From the expressions for every object ID, the most frequent word was chosen as the single word most likely to describe the object in the image. This required that at least two descriptions of the image mentioned the object by the same term, otherwise the object was excluded. The word selection was done after omission of very frequent, irrelevant words such as articles ('a', 'an', 'the'), third-person possessive determiners ('his', 'her', 'their'), the cardinal numbers one through ten, and primary and secondary colors (e.g. 'orange'), including 'silver' and 'gold'. If multiple objects in an image had been labeled with the same word, only one of them was selected (the first one in the loop, not randomly). Finally, images were removed that contained fewer than two objects after preprocessing.

This yielded a total of 86,748 word-object pairs, resulting in a vocabulary of 4,237 unique words. It should be noted that objects paired to the same word could still display great visual variability. The least frequent words (e.g. 'beak' and 'paste') occurred only once, whereas the most frequent word occurred 7,891 times. The five most frequent words were *man* (7,891 times), *shirt* (4,536 times), *woman* (4,378 times), *boy* (1,477 times), and *girl* (1,428 times). The average frequency was 20.47 (*SD* = 172.33), and the median frequency was 2. After preprocessing, 24,670 images remained, of which 1,000 were set aside as validation data, and another 1,000 as test data.

Table 1: Number of objects and baselines per split.

Split	Objects	Listener baseline	Speaker baseline
Train	79,749	0.284	0.091
Test	3,493	0.286	0.089

Table 1 shows the total number of objects in the train and test splits of the data, as well as the baselines for the listener and speaker respectively. The listener baseline is one divided by the average number of objects per scene. The speaker baseline is the majority baseline of the most frequent word. The baselines represent the average accuracy obtainable by chance, which serves as the minimal performance expected of the model. High accuracy is only an indication of good performance if the model performs better than its baseline. Since many words occur only once or twice, there are 80 words in the test set that do not occur in the training set, with a token frequency of 80, and 776 words, with a token frequency of 3413 in the test set, that do occur in the training set. These numbers might suppress test accuracy.

Curiosity Mechanism

In order to measure the effect of active and curious learning, the model which performed curiosity-based object selection was compared to a model that received the next object to learn about randomly. In the first condition, curiosity values were calculated for each object using T&W's curiosity mechanism, and the object with the highest value was chosen to learn about. In the second condition, objects were randomly chosen from the scenes. The main purpose of the speaker part of the model was to produce a word guess as input to the listener so that the model could run without the input provided by the oracle. This way, the model could run (without weight updates) to compute curiosity values and

choose the most interesting object to talk about, before running (with weight updates) to learn about the form-meaning pair with feedback from the oracle.

T&W's curiosity mechanism (see equation (1)) was used to produce the curiosity values, where *i* was the object representation given as input to the speaker, and *o* was the object prediction produced by the listener. The curiosity values were computed element-wise, and the mean of the absolute values of the curiosity vector was taken as the curiosity value for an object in the scene. The object with the highest curiosity value was chosen as the next input for the speaker and target for the listener.

The random and curious conditions were compared on listener loss and accuracy, which indicate the models' ability to choose the appropriate referent of a word form. The loss and accuracy patterns produced over the 40 epochs were plotted to be interpreted as learning curves and compared between conditions.

Results

Figure 3 shows the value of the loss and accuracy of the listener, after each epoch of training. Curious listeners (the blue lines in all plots) show a consistent pattern: after one epoch of training, accuracy on the test set ranged from .49 to .61, far above the baseline of .286. The accuracy on the test set steeply increased in the first few epochs, and kept increasing more slowly, but steadily over later epochs, converging somewhere around epoch 20 with accuracy from .71 to .74. At epoch 40, accuracy ranged from .72 to .75. The exception to this pattern is one particular run, which shows a similar learning trajectory but started and ended with a much lower accuracy, of .32 and .58, respectively. The general pattern is reflected in the plots of the loss on the test data.

On the training data, accuracy also plateaued around epoch 20, with accuracy from .80 to .83 for 19 runs, and only small gains in accuracy until epoch 40, with scores from .83 to .84. Note that the training loss continued to decrease after epoch 20. This indicates the curious listeners started to overfit at that point, fitting to specific characteristics of the training set, that did not translate to accuracy or improvements on the test data. As we saw on the test data, one run shows a different pattern and reaches a maximum of .76 in accuracy on the training data.

The pattern for listeners in the random condition (the orange lines in all plots) is more erratic. After one epoch of training, all random listeners started around or just above the baseline accuracy of .286. Some listeners in this condition barely outperformed the baseline at epoch 40. Others outperformed the baseline, but plateaued after 10-20 epochs, eventually reaching maximum accuracy scores ranging from .39 to .48 on test data. For 6 runs, the accuracy after epoch 1 was around the baseline, but increased steeply until epoch 20, and continued to increase slowly after that. At

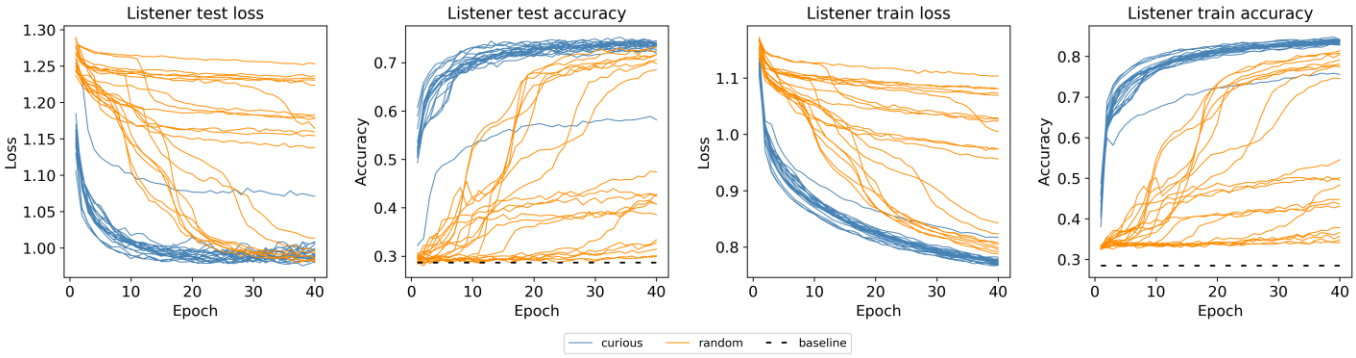

Figure 3. Test and train results of the listener.

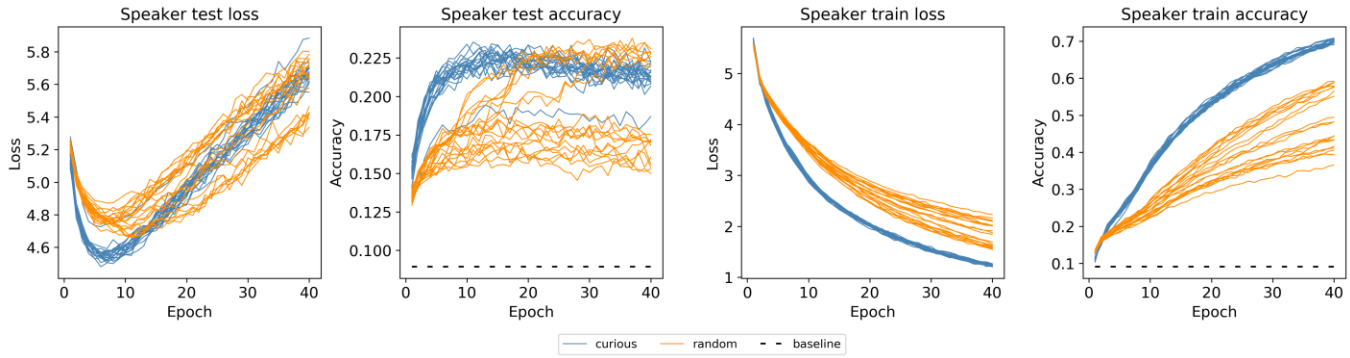

Figure 4. Test and train results of the speaker.

epoch 40, performance of 8 runs is slightly below that of most runs in the curious condition, with test accuracy ranging from .68 to .73, and train accuracy from .78 to .81. The same patterns are reflected in the loss plots.

Test accuracy of speakers trained in the curious condition peaked somewhere between epoch 10 and 25 around .23, with the exception of the one run in which the listener was also less successful, which peaked at epoch 12, with an accuracy of .19. The loss value was lowest around epoch 8. After this epoch, the training loss was still consistently going down, and training accuracy going up. After epoch 8-10, the curious speakers were overfitting rather than learning.

As with the listeners, initially, speakers in the random condition learned more slowly, as is reflected in the lower accuracy between epochs 1 and 20. In all random runs, the speaker outperformed the baseline. However, as was the case with the listeners in this condition, there are large differences between runs. Most runs plateaued relatively quickly, and peaked between .16 and .19, whereas in 8 of the 20 runs, accuracy continued to increase, eventually matching performance of the speakers in the curious condition, with accuracy peaking around .23. Although the training trajectories in the random condition are more discernable than for the curious condition, in all runs, performance on the training data continued to improve until epoch 40. As in the curious condition, all random speakers overfitted.

Discussion

Did curiosity increase the performance of the word learning model compared to the random choice of objects? Yes, the listener test loss decreased faster and the listener test accuracy increased faster in the curious condition than in the random condition. Whereas the curious model converged at a similar point on every run, the random model eventually equaled or approached the curious model on some runs, but learned nothing or was stuck in a local optimum on others.

A pattern that can be discerned is that curiosity, aggregated over the different initializations, performs better from the start and learns faster than random selection. In this experiment, the random initialization of the weights meant that the first objects selected in the curious condition were just as random as those in the random condition. This changed after a few weight updates when the curiosity formula took effect—the difference in performance becoming apparent after a single epoch. This behavior is different from what is typically proposed, as intrinsic motivation is expected to make learning slower initially, but make up for that with increased performance and better generalization in the long run (Oudeyer & Kaplan, 2007).

Another pattern that can be observed is that learning trajectories of curious learners were more similar to each other than those of random learners were. Curiosity seems to provide 'robustness', making learners less prone to being stuck in a local optimum.

The near instant performance advantage of curiosity may be explained by the inherent advantage it has over random selection when dealing with token frequency. Having a good word representation for the corresponding object brings an increase in overall accuracy equivalent to its token frequency. Whereas random selection is prone to select objects with a high token frequency, curious selection can focus on highly frequent word-object pairs first, and ignore them later once their representation is already accurate. Further research could establish whether the selection by the curiosity mechanism matches this strategy.

This would correspond to the notion that language is not a product, but a means for social interaction, where the child's initial interest is to get the message across and language learning follows (Clark, 2016). The intentionality theory of language learning describes how such intrinsically motivated behavior can drive language learning (Bloom, 2000). As of yet, there is no empirical data on what criteria or strategies children use to pick topics to talk about.

Whereas the model was evaluated on the listener performance (comprehension), the speaker's main purpose was to enable the curiosity mechanism, which was used to train the curious model. The high train accuracy of the curious speaker increased the accuracy of the curiosity mechanism, thereby improving the curious listener's comprehension. However, the speaker overfitted in both conditions, and did not generalize well to test data. The speaker test results therefore do not help to understand how improved comprehension leads to improved language production.

We have shown that modeling the language learner as an active solicitor of input, rather than a passive receiver, can lead to different learning outcomes. When objects in the context are selected as a topic according to curiosity, word learning is faster and more robust than when topics are selected at random. Future work may explore the distributional properties of the topics selected by curiosity over the course of the learning process.

References

Bloom, L., Margulis, C., Tinker, E., & Fujita, N. (1996). Early conversations and word learning: Contributions from child and adult. *Child Development*, 67(6), 3154-3175.

Bloom, L. (2000). The Intentionality of Word Learning: How to Learn a Word, Any Word. In Golinkoff, R., Hirsh-Pasek, K., Bloom. L., Smith, L., Woodward, A., Akhtar, N., Tomasello, M., & Hollich, G. (2000), *Becoming a word learner: A debate on lexical acquisition*, 19-50. NY: Oxford University Press.

Chapman, R., Miller, J., MacKenzie, H., & Bedrosian, J. (1981). The development of discourse skills in the second year of life. *Second International Congress for the Study of Child Language*, Vancouver, BC.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3), 637-669.

Clark, E. V. (2016). *First language acquisition*. Cambridge University Press.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mather, E. (2013). Novelty, attention, and challenges for developmental psychology. *Frontiers in psychology*, 4, 491.

Oudeyer, P. Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in neurorobotics*, 1, 6.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017). Automatic differentiation in PyTorch. *NIPS 2017 Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *International Conference on Machine Learning (ICML)*.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Proceedings of the IEEE international conference on computer vision* (pp. 2641-2649).

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2017). Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision* (pp. 817-834). Springer, Cham.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1), 54-67.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230-247.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556v6*.

Twomey, K. E., & Westermann, G. (2018). Curiosity-based learning in infants: a neurocomputational approach. *Developmental Science*, 21(4), e12629.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 67-78.