Active Word Learning through Self-supervision

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

Alireza Mahmoudi Kamelabad[1] **(a.m.kamelabad@gmail.com)**
CIMeC - Center for Mind/Brain Sciences
University of Trento

Afra Alishahi (a.alishahi@uvt.nl)
Department of Cognitive Science and Artificial Intelligence
Tilburg University

Abstract

Models of cross-situational word learning typically characterize the learner as a passive observer, but a language learning child can actively participate in verbal and non-verbal communication. We present a computational study of cross-situational word learning to investigate whether a curious word learner who actively influences linguistic input in each context has an advantage over a passive learner. Our computational model learns to map words to objects in real images by self-supervision through simulating both word comprehension and production. We examine different curiosity measures as guiding input selection, and analyze the relative impact of each method. Our results suggest that active learning leads to higher overall performance, and a formulation of curiosity which relies both on subjective novelty and plasticity yields the best performance and learning stability.

Keywords: Cross-situational word learning; Computational modelling; Active learning; Curiosity.

Introduction

An important task in language acquisition is learning which words refer to which objects in the world. *Cross-situational word learning* is the process by which learners match words to objects by tracking word-object co-occurrences over many instances. Often when a word is encountered, there are many candidate objects in the context that the word might refer to. However, when one takes into account multiple occurrences of the word, some object(s) will be consistently present, making them more likely candidates.

Models of cross-situational word learning typically characterize the learner as a passive observer. However, a language learning child can actively participate in verbal and non-verbal communication. Thereby, they may be actively shaping their linguistic input. Bloom, Margulis, Tinker, and Fujita (1996) found that in early child-caregiver interaction, children often introduce a new topic into the conversation, and parents are likely to follow up by continuing to talk about the same topic. Another process through which children may shape their own language input is *joint attention*: certain objects are the focus of attention of both participants in a conversation. Caregivers are sensitive and responsive to children's attention in several ways, and talk about objects when they are the focus of attention (Chang, de Barbaro, & Deák, 2016). Several studies have found that caregivers *following* the child's attention (as opposed to drawing a child's attention towards a certain object), and talking about objects that are already in focus, is correlated with word learning (Akhtar, Dunham, & Dunham, 1991; Tomasello & Farrar, 1986).

In this work, we consider the possibility that language learners use their active role in communication to elicit the most informative linguistic input from their interlocutors. In several other domains, children attend to or sample informative data. When searching for rewards, children, more than adults, explore uncertain options (Schulz, Wu, Ruggeri, & Meder, 2019). Kidd, Piantadosi, and Aslin (2012, 2014) show a *Goldilocks* effect in infants' attention to visual and auditory stimuli. Infants attend especially to stimuli that are somewhat complex, but not too complex; somewhat predictable, but not entirely, according to their current knowledge state. We propose that children may also display such curiosity-driven behaviour during language acquisition, and study the potential effects of an active input selection mechanism on word learning in a computational setup.

In artificial intelligence research, different implementations of active learning have been proposed and studied, including definitions based on novelty, predictability, or task success in the long run (see Oudeyer and Kaplan (2009) for an overview). In a cognitively motivated study, Twomey and Westermann (2018) show that a model of visual category learning achieves maximal learning results when it selects its input according to a curiosity metric that balances the novelty of the stimulus and potential knowledge update. This concept was applied to the task of cross-situational word learning in a computational study of Keijser, Gelderloos, and Alishahi (2019). They show that an agent that curiously selects objects to receive linguistic input for eventually learns word-object mappings more accurately and robustly. However, in their study, learning to understand which word maps to which object is a supervised process – an unrealistic assumption with respect to the human language acquisition process. Also, they only investigate a single formulation of curiosity, and it is not clear which component of this formulation yields the gain observed in their simulation results.

[1] Project carried out while visiting the Department of Cognitive Science and Artificial Intelligence at Tilburg University.

Inspired by Keijser et al. (2019), in this paper we use a computational study of cross-situational word learning to investigate whether a curious word learner who actively selects the linguistic input in each context has an advantage over a passive learner. In our study,

- we simulate word learning as a self-supervised process, where instead of relying on corrective feedback on labels of objects or referents of words from the environment, the model relies on consistency when applying its own acquired linguistic knowledge to both word comprehension and production tasks;
- we use real images as visual context, and use pre-identified objects and their annotated labels as learning material to our computational model;
- we examine different curiosity measures as guiding input selection, and analyze the relative impact of each method on the overall performance and stability of the word learning model.

Our results suggest that a curious learner who actively influences linguistic input has an advantage over a passive learner. Furthermore, a formulation of curiosity which relies both on subjective novelty and plasticity yields the best performance and learning stability, compared to using each of these factors alone.

Method

Task

We operationalize word reference learning as a dual process, involving both *comprehension*, or understanding which object a word refers to, and *production*, or the ability to use a word to describe a given object. The task then is twofold: for the receptive part of the agent, the goal is to be able to identify the true referent of a word among a set of candidate objects; and for the productive part, the goal is to refer to an object by the appropriate word.

Comprehension The learner receives as input a set of objects from a visual scene, and a word. Objects are vectors carrying high-level visual information, each representing one patch of the image processed by a pretrained object recognition model. The task is to determine which of the objects is the best referent for the input word.

Production The learner receives one visual object as input, and its task is to determine which word is the best label to refer to this object. As in comprehension, the object is represented as a visual vector. Unlike in comprehension, the production module does not take into account the rest of the visual scene.[2]

[2]In reality, the most effective word or expression to describe an object depends not just on the object in question, but also on what other objects are present in the scene.

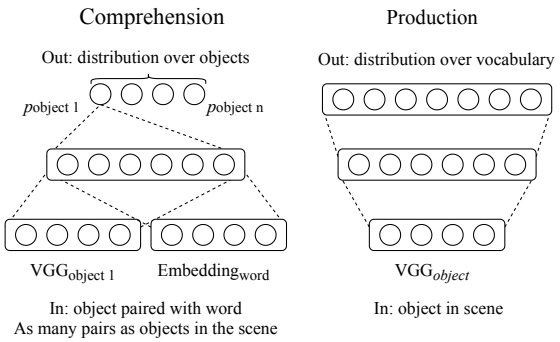

Figure 1: Architecture of the model, adapted from Keijser et al. (2019).

Data

The data consist of images with multiple objects in them, and a word associated with each object. We use the dataset that was introduced by Keijser et al. (2019). It is based on the Flickr30 Entities corpus (Plummer et al., 2015), which contains annotations for the bounding boxes for referring expressions in the captions of the Flickr30 dataset (Young, Lai, Hodosh, & Hockenmaier, 2014).

Keijser et al. (2019) filtered out any expressions that describe multiple bounding boxes. The referring expressions are simplified by selecting only the most frequent word in the expressions as referent per object. Objects therefore were only included if at least two referring expressions contained the same word. Several heuristics were used to select those words that denote the object. Cardinal numbers one through ten as well as colour names were filtered out. Very frequent, but visually irrelevant words were also filtered out, such as articles and possessives. The data was simplified by keeping only one object per word per image, thereby removing ambiguity. Only those images were included for which after filtering there were still at least two objects. The dataset contains 24,670 images, 4237 unique words, and 86,748 word-object pairs. 1000 images were used as validation data and another 1000 for testing. A full description of the dataset can be found in Keijser et al. (2019).

Model

The model consists of a production and a comprehension module. The architecture is sketched in Figure 1. The output of each module can serve as input to the other. This allows for introspection: when the production module outputs a word, the comprehension module can in turn try to interpret it; and vice versa, when the comprehension module selects an object as referent for a word, the production module can try to name the object. The learner can check if it understands its own language production. It is this introspective property of the model that makes learning under self-supervision possible.

Comprehension The comprehension module learns to map a given word to its referent in the visual scene consisting of a number of objects. We represent objects by visual feature vectors extracted from the last fully connected VGG-16 object recognition model (Simonyan & Zisserman, 2015), pretrained on the ImageNet dataset (Deng et al., 2009). Words are represented by 256-dimensional vectors (or embeddings), which are learned during training. The candidacy of every object in a scene as the referent of a given word is considered in parallel: when the module receives a word as input, it concatenates the word embedding to the visual feature of every object in the scene separately. This concatenation of word embedding and a single object representation is input to 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. The object with the highest output value is most likely to be the referent.

Production The production module learns to output a word, given an object. The input is again in the form of a VGG vector, which is fed to a 256 unit hidden layer followed by sigmoid activation, which is fully connected to the vocabulary-sized output layer: every unit representing a word. The word unit with the highest output value is the best candidate to describe the target object.

Self-supervised learning Although the model consists of two modules that can be used independently, the whole agent is trained in one go. The setup is inspired by Rohrbach, Rohrbach, Hu, Darrell, and Schiele (2016) which use a similar setup for training a computer vision model to ground referring expressions in the visual scene. During training, once a word is processed by the comprehension module, we use the softmaxed output vector as attention over the objects in the scene. Input to the production module consists of the sum of the visual feature vectors of all the objects in the scene, weighted by the output of the comprehension module. The whole agent, including the production and the comprehension module, can now be updated in one go, by comparing the output values of the production module with the word that was input to the comprehension module in the first place.

Details of the implementation The model was implemented in PyTorch (Paszke et al., 2019). It was updated according to the cross-entropy between the one-hot encoding of the input word and the output of the production module. The model was trained using Adam optimization (Kingma & Ba, 2015) in minibatches of 40 instances for 40 epochs. To decide on an initial learning rate, we ran the model in all conditions with a learning rate from .1 to .00001 for 40 epochs. A learning rate of .001 yielded the best results on validation data for both the comprehension and production modules and was used for training all models reported in the results section. We ran 20 different initializations per condition.[3]

[3]The implementation itself as well as code used to report results and further analysis is available on Github: `https://github.com/horotat/curiosity`

Input selection

In any given environment, a language learning child has the choice to direct their attention to whatever object in their vicinity is most interesting to them. On top of that, they have influence over the environment itself – by moving to a different location or manipulating objects. Although our learner has no influence over its environment (it can not choose which image it sees), it can choose for which object in the scene it receives language input. We compare a learner that receives language input for a randomly drawn object to learners that select an object to receive input about according to an estimate of learning potential. Importantly, since all learners see each image once per epoch, all learners have access to the same number of data points.

The input selection is an introspective process: to select an object, the learner inspects its knowledge of all the objects in the scene. For every object, first the production module tries to find the corresponding word. This word is then fed as input to the comprehension module, which in turn tries to find the corresponding object. The contrast between the object input to the production module and the distribution over the objects output by the comprehension module is the basis for estimating which object holds most learning potential.

There are different ways to estimate learning potential. Approaches in reinforcement learning often include a definition of expected reward and cannot be applied easily to our learning problem. The metrics we use for input selection are based on Twomey and Westermann (2018): subjective novelty, plasticity, and curiosity. These metrics were defined for category learning and translate well to our set-up. Subjective novelty favours the most unknown objects, plasticity selects on how much the learner expects to learn from a given input word, and curiosity is the product of those two. Each of them is calculated for every object in the scene. The object that maximizes the metric is selected to receive linguistic input for.

Subjective novelty is defined in equation 1. t is the true distribution over objects; that is, it is a one-hot vector encoding the object we are calculating subjective novelty for. o is the guessed output by the comprehension module, and n is the number of objects in the scene. Subjective novelty, then, is the average absolute difference between the comprehension module's guesses and the true distribution. Intuitively, subjective novelty selects the object for which the learner expects to be most wrong, either because it is misnamed in production, or because the produced word is misinterpreted in comprehension.

$$s(t,o) = \frac{\sum_{i=1}^{n} (|t_i - o_i|))}{n} \tag{1}$$

Plasticity is defined in equation 2. Because every element of o is the result of a sigmoid function, $o_i(1-o_i)$ is the derivative of o. This is the value on which model updates are based.

Intuitively, the larger plasticity is, the more effective an update to the model will be, in the sense that the comprehension modules guesses will change a lot.

$$p(o) = \frac{\sum_{i=1}^{n} o_i (1 - o_i)}{n} \quad (2)$$

Curiosity is defined in equation 3. It is the product of subjective novelty and plasticity, averaged over the objects in the scene. Curiosity balances plasticity and subjective novelty, favoring objects for which both are high.

$$c(t, o) = \frac{\sum_{i=1}^{n} (|t_i - o_i|)\, o_i (1 - o_i)}{n} \quad (3)$$

Results

Table 1 shows the accuracy of trained models in each condition on held-out test data. For testing, we do not use the input selection mechanisms, but rather test every single word/object in every image. Therefore, these scores are comparable across models trained using different input selection mechanisms. We trained 20 models in every condition, each starting from a different randomly initialized state. For every run, we select the model after the epoch with maximum validation scores (selection was separate for production and comprehension). We report the average accuracy and standard deviation over all runs in every condition. We also report a baseline for both comprehension and production. For comprehension, this is the score a random guesser would obtain. For production, it is the accuracy when always guessing the most frequent word. Please note that the production task is considerably more difficult than the comprehension task. The difference is more extreme than the baseline gives away; whereas the comprehension module has to decide between only 2 to 10 objects in the scene, the production module always has as many options as there are words in the vocabulary.

When we look at the accuracy scores for the comprehension module, we see that models trained with input selection according to curiosity ultimately attain the highest accuracy scores. However, neither models trained with plasticity nor subjective novelty as selection mechanism outperform models trained with random object selection. In fact, models trained using subjective novelty perform at chance level. When we look at the standard deviations for the conditions in which learning is successful, we see that this is lowest for the curiosity condition, meaning that the differently initialized models reach more comparable scores in this condition.

The general pattern of scores in comprehension is also reflected in the scores for the production module, with the exception that here, models trained using subjective novelty for input selection do beat the baseline by some margin, although they still score the lowest by far. Here, too, curiosity is the highest accuracy condition and also has the smallest standard deviation, whereas neither plasticity nor subjective novelty outperform models trained with random input selection.

Table 1: Average accuracy and standard deviation on test data

	Comprehension		Production	
	Acc.	SD	Acc.	SD
Random	.5458	.0746	.2093	.0139
Plasticity	.5119	.1035	.1801	.0223
Subjective novelty	.2874	.0026	.1214	.0082
Curiosity	**.6626**	.0190	**.2132**	.0046
Baseline	.2863		.0893	

In order to understand these results, we also look at the intermediate scores during training of all models on training and test data. Note that all models see exactly the same objects in the test setting, namely all objects in the test set, but in the training setting, they only see one object per image that was selected according to their input selection method. The scores on test data are therefore more comparable than those on training data.

The results during training are visualized in Figures 2 and 3. Every line represents a single model's training trajectory. Comprehension scores of models in the plasticity, curious, and random conditions show the general pattern we might expect when looking at the averaged data. Models in the curious condition immediately outperform those in other conditions, with models in the random condition gradually catching up to some extent. The trajectories of different models in the curiosity condition lie close together, whereas those in the random and plasticity condition are more spread out. The high training scores for many models in the subjective novelty condition, but baseline performance on the test set, indicate that these models are prone to overfitting.

When we look at the production scores in Figure 3, we see that all models are prone to overfitting on this task. Nevertheless, we also see clear differences in test scores between the conditions. As in the comprehension task, models trained using curious input selection immediately outperform models in other conditions. More than in the comprehension task, models receiving random input eventually catch up, although there is more variance amongst models in this condition than amongst models trained using curiosity. Models in the subjective novelty condition are again especially prone to overfitting; initially there is some learning that generalizes to the test set, but this knowledge is gradually forgotten as the models tune to the specifics of the training data.

Analysis of input selection metrics

As we saw in the last section, selecting input based on curiosity improves performance in both the comprehension and production tasks, but its components, plasticity and subjective novelty, are outperformed by random selection. In order to understand how curiosity can help learning when its moving parts by themselves do not, we take a closer look at their object selection behaviour. One may expect that curiosity, being the product of subjective novelty and plasticity, some-

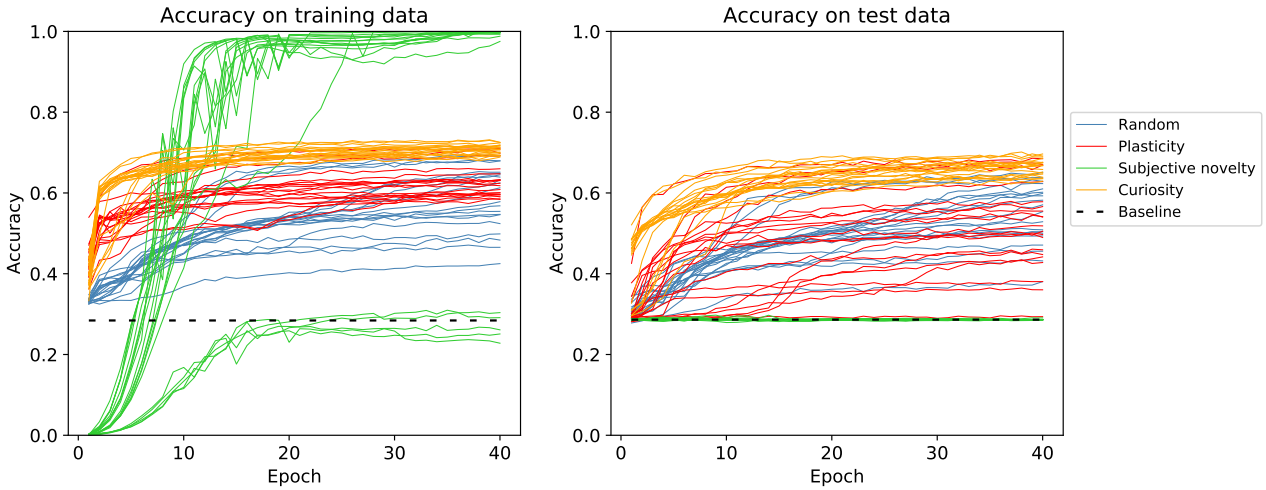

Figure 2: Comprehension train and test accuracy during training. Every line represents the training trajectory of a single model.

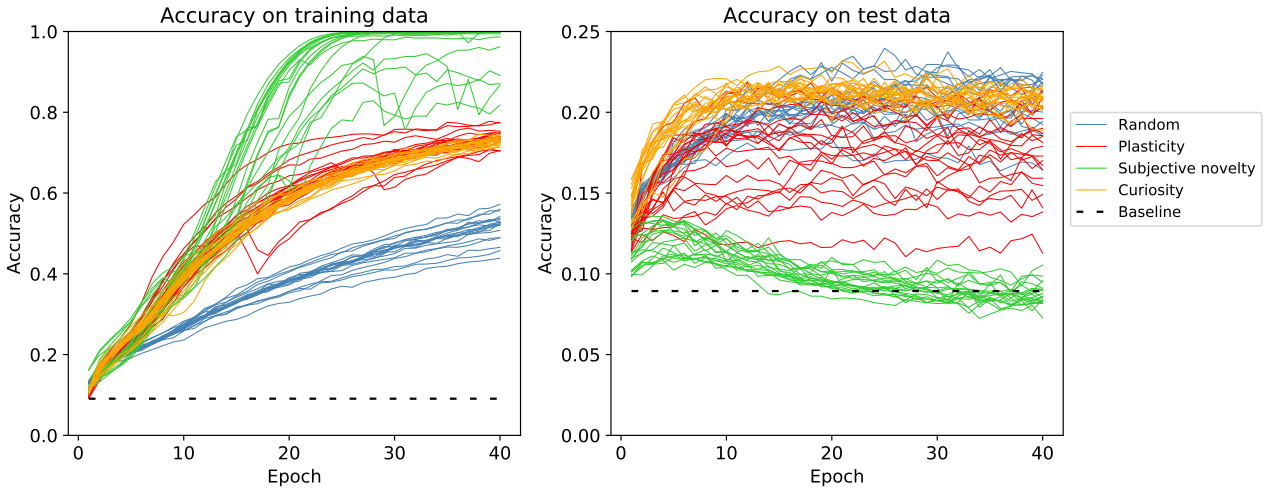

Figure 3: Production train and test accuracy during training. Please note that the y-axis for test accuracy is scaled for visibility.

times mimics one, and sometimes the other. However, because neither of the two components of curiosity yield comparable results by themselves, we expect that curiosity may make its own choices altogether. In images with three objects *a*, *b*, and *c*, where *a* has maximal plasticity and *b* has maximal subjective novelty, perhaps curiosity, rather than choosing object *a* or *b*, selects object *c*.

To analyze the overlap in selection behaviour between the three mechanisms, we select the optimal input object according to all three selection mechanisms for all 1,000 validation images, and then do a pairwise comparison between the selections. The analysis is done before any training and after each training epoch. Since models in the input selection conditions are necessarily shaped by the selection mechanism with which they were trained, we decide to do this analysis on models trained in the random condition. The results of this analysis are illustrated in Figure 4. Each line represents the number of images for which two, or all three mechanisms made the same choice, averaged over all 20 models in the random condition.

Before any training (visualized as 'epoch 0') curiosity and subjective novelty highly overlap, on average choosing the same object in 953.55 out of 1,000 images, whereas plasticity often selects a different object, overlapping with curiosity in 336.35 images on average. In epochs after training, it is plasticity and curiosity that overlap (steadily growing from 571.7 after epoch 1 to 747.85 after epoch 40). There is comparatively little overlap between curiosity and subjective novelty, at an average 267.9 after epoch 1, slightly decreasing until 220.2 after epoch 3, then again growing steadily until 370.45 after epoch 40. In many cases after training, when curiosity and subjective novelty overlap, there is overlap between

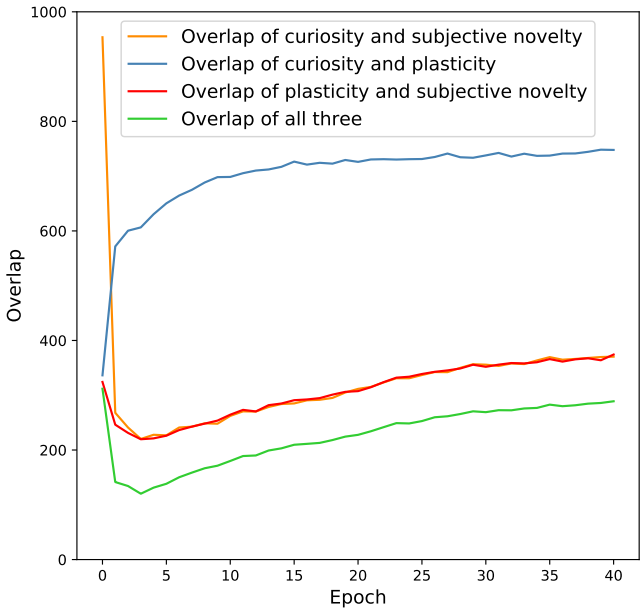

Figure 4: Overlapping choices of the different selection mechanisms for images in the validation set during training. Overlap scores are averaged over the 20 models in the random condition.

all three cases. The overlap between curiosity and subjective novelty, but not plasticity, comprises 126.35 images after epoch 1, and this proportion decreases until epoch 8, where it stabilizes around 80 out of 1,000 images.

After 1 epoch of training, curiosity chooses a different object than both plasticity and subjective novelty in 301.95 images on average. The proportion of unique choices by curiosity gradually decreases until 170.7 after epoch 40.

In summary, before any training, curiosity mainly agrees with subjective novelty. In the early stages of training, curiosity usually aligns with plasticity, and in a large minority of cases chooses distinctly from both plasticity and subjective novelty. After training, the overlap between curiosity and subjective novelty alone is small, particularly from epoch 8 on. The proportion of unique choices by curiosity becomes smaller over training, while agreement between curiosity and plasticity grows.

Discussion

We simulated word learning as a self-supervised process. The learner learns to comprehend and produce words through self-supervision. Using the introspective quality of the model, we studied the effect of active input selection on the word learning process. Just as Keijser et al. (2019) have shown in a supervised word learning task, we find that curious input selection leads to better performance, faster learning, and more robust convergence, as compared to random input. In addition, we also find that neither plasticity nor subjective novelty by itself leads to similar improvements. In fact, random input selection outperforms subjective novelty and plasticity. Input selection based on subjective novelty is very prone to overfitting. A possible explanation is that by selecting objects it expects to be wrong about, subjective novelty is likely to select 'exceptions to the rule' and leads models to fit to idiosyncracies of the training set, rather than to generalizable knowledge. Although the learning trajectory of models trained under plasticity looks more qualitatively similar to that of models in the random or curiosity condition, eventual performance is lower.

Since curiosity is a function of plasticity and subjective novelty, it is somewhat surprising that neither of these mechanisms by themselves yield comparable advantages to random input selection. We analyzed the overlap between all three mechanisms. For a completely untrained model, selection according to curiosity shows near complete overlap with selection according to subjective novelty, but this overlap quickly disappears. Although curiosity and plasticity show considerable overlap during training, there is still a significant portion of images where curiosity selects a different object than both plasticity and novelty, particularly in the early epochs. Curiosity, then, seems to be doing more than simply balancing the selection of subjective novelty and plasticity.

Of course, this work only shows what effect active selection of input might have on learning trajectories, *when* learners use active input selection strategies. We have no empirical evidence that, in fact, they do. This work should be seen as a proof of concept; *if* we find that learners actively solicit input according to a certain definition of curiosity, then this can influence their learning, and models of word learning should take it into account. Whether learners do employ such a strategy, must be established in empirical research. Lab studies can give insight into what learners do when they can explicitly control their input (Kachergis, Yu, & Shiffrin, 2013). However, to understand whether this is a natural part of the language acquisition process, it is necessary to study children's introduction of topics in child-caregiver interaction.

References

Akhtar, N., Dunham, F., & Dunham, P. J. (1991). Directive interactions and early vocabulary development: The role of joint attentional focus. *Journal of Child Language*, *18*(1), 41–49.

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.

Chang, L., de Barbaro, K., & Deák, G. (2016). Contingencies between infants' gaze, vocal, and manual actions and mothers' object-naming: Longitudinal changes from 4 to 9 months. *Developmental Neuropsychology*, *41*(5-8), 342–361.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image

database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Kachergis, G., Yu, C., & Shiffrin, R. M. (2013). Actively learning object names across ambiguous situations. *Topics in Cognitive Science*, *5*(1), 200-213.

Keijser, D., Gelderloos, L., & Alishahi, A. (2019). Curious topics: A curiosity-based model of first language word learning. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 1991–1997). Cognitive Science Society.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLOS ONE*, *7*(5). doi: 10.1371/journal.pone.0036399

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child Development*, *85*(5), 1795-1804.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations.*

Oudeyer, P.-Y., & Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *1*. doi: 10.3389/neuro.12.006.2007

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035).

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. In *The IEEE International Conference on Computer Vision (ICCV)* (pp. 2641–2649).

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2016). Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision* (pp. 817–834).

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *Psychological Science*, *30*(11), 1561-1572.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations.*

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, *57*(6), 1454–1463.

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

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.