Modeling Language Learning in Corrective Feedback Interactions

Juan L. Castro-Garcia
Computer Science and Engineering
Michigan State University
castrog4@msu.edu

Parisa Kordjamshidi
Computer Science and Engineering
Michigan State University
kordjams@msu.edu

Abstract

To study computational models for language acquisition, we propose an interactive computational framework that utilizes a miniature language acquisition dataset in a controlled environment. In this framework, a neural learner model interacts with a teacher model that provides corrective feedback. Within this framework, we investigate various corrective feedback strategies, specifically focusing on reformulations and their effect on the learner model during their interactions. We design experimental settings to evaluate the learner's production of syntactically and semantically correct linguistic utterances and perception of concepts and word-meaning associations. These results offer insights into the effectiveness of different feedback strategies in language acquisition using artificial neural networks. The outcome of this research is establishing a framework with a dataset for the systematic evaluation of various aspects of language acquisition in a controlled environment.

1 Introduction

Understanding how children form associations between linguistic words to some situational input or referent within an uncertain environment where multiple referents could be related to the same word is a topic that has been studied in language acquisition (Quine, 1960). Cross-situational learning is a powerful mechanism for learning co-occurrence statistics between words and referent objects across multiple exposures (Gleitman, 1990; Pinker, 2013). Studies of cross-situational learning in both adults (Yu and Smith, 2007; Smith et al., 2011) and children (Suanda et al., 2014; Smith and Yu, 2008) show how the association between words and meaning is learned at different stages of language development. However, some of these studies on cross-situational learning focus on the child's learning of these word-meaning associations without any form of feedback (Monaghan et al., 2021).

Feedback, in the form of social interactions, is shown to enhance children's language development (Kuhl et al., 2003; Sachs et al., 1981; Krashen et al., 1983). In language acquisition studies, most commonly in second language acquisition literature, an interaction is viewed as a negotiation for meaning where two agents "negotiate" or agree upon the meaning of some object during a conversation (Long, 1981; Clark, 1996). Corrective Feedback is one form of interaction where an adult, i.e. parent or teacher, analyzes the linguistic generation of a child and provides some form of response intended to adjust or update the child's linguistic knowledge. Although the impact of providing corrective feedback is a controversial topic, many studies supports its influence on language learning even in first language acquisition (Hiller, 2016; Chouinard and Clark, 2003; Schoneberger, 2010). In a social context, commonly within a classroom setting, several approaches for corrective feedback are utilized such as explicit correction, recast or reformulations, clarification request, metalinguistic feedback, elicitation, and repetition (Lyster and Ranta, 1997). For the scope of this paper, we will focus on reformulations as our computational approach to corrective feedback.

Cross-situational learning has been used to address multiple tasks like probabilistic word-meaning learning with symbolic situation representations (Fazly et al., 2010), word-meaning learning with embodied systems (Yu and Ballard, 2004), and word-meaning associations from visual perceptual representations as inputs (Juven and Hinaut, 2020). Also, several studies have explored learning settings that simulate interactions between teacher and learner conversations with k-Nearest Neighbor models for word learning (Belpaeme and Morse, 2012) and probabilistic models with corrections (Angluin and Becerra-Bonache, 2017). Other models have studied the acquisition of semantic knowledge through some form of feedback provided via a reward function in a reinforcement learning setting (Nikolaus and Fourtassi, 2021b,a). Other studies have explored the language acquisition process as a game where two agents observe referents in a scene and both attempt to name it (Steels, 1995).

Building neural computational models for Language Acquisition based has been practiced in the related literature (Portelance and Jasbi, 2023; Frank et al., 2019). However, one issue that these models present is the requirement of large amount of linguistic data, most times larger than what humans are exposed to through out their entire lifetime. Also, the architecture design of these models can have innate biases that makes relating their outcome analysis to human language acquisition theories challenging (Baroni, 2022). In order to

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 267–279

establish a fair and accurate comparison between neural models and human learners, it is required to simulate learning scenarios where the quantity of data, input modalities, and data distributions resemble human-level abilities (Warstadt and Bowman, 2024).

In this paper, we use a learning scenario that explores neural computational models for language acquisition in a teacher-learner interactive cross-situational learning framework. This framework follows a similar structure as the one presented in Angluin and Becerra-Bonache (2017) which uses a Miniature Language Acquisition dataset where the environment contains logical representation of objects and the model learns to generate linguistic utterances that describe these objects. We extend their work by using recurrent neural networks as the learner rather than a probabilistic graphical model. We train small models from scratch, without any prior knowledge, that resembles early stages of concept learning and language acquisition.

Also, we examine various corrective feedback strategies and their impact on the learner model's learning trajectory. The learner is evaluated at a production level based on utterance semantics and its ability to generate all possible explanations, as well as a perception level, referring to the knowledge acquired about concepts and relationships. Figure 1 shows how interaction is established, where each interaction starts with a learner model generating syntactically and semantically appropriate utterances corresponding to the attributes and relationships within a given situation, represented formally in a formal predicate-argument form. The teacher then analyzes the utterance, compares the utterance's formal representation to the situation's formal representation, and provides another utterance with a similar formal representation to the situation.

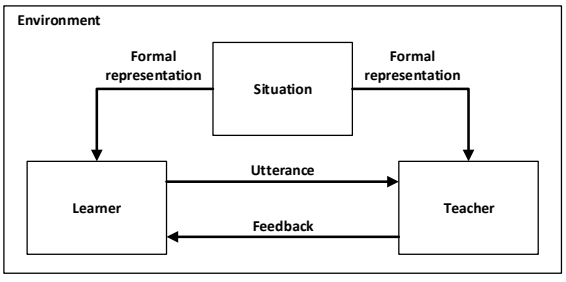


Figure 1: Teacher-Learner Interaction model. This framework has a teacher and learner model interact in an environment with shared situations. The learner model attempts to describe the situation, and then the teacher evaluates the description and provides feedback to the learner.

The main contributions of this work are as follows: 1) Extending Miniature Language Acquisition computational framework with neural learner models; 2) Providing evaluation datasets and metrics for utterance generation (production-level) and concept evaluation (perception-level); 3) Incorporating various corrective feedback strategies in the form of interactions with an oracle that evaluates the logical semantics.

2 Miniature Language Acquisition Dataset

Miniature Language Acquisition (MiLA) is a task that consists on the learning of a natural language from sentence-picture pairs, where each "picture" or scene is composed of geometric shapes with different properties (Feldman et al., 1990). Similar to the setup in (Angluin and Becerra-Bonache, 2017), we create a learning setting in which the learner receives a formal representation of the environment, instead of actual visual input, and generates natural language utterances that explain the environment. In our experiments, we only use English natural language utterances. The dataset includes pairs of formal representations of various situations with their respective set of valid linguistic utterances. There are 23,328 unique situations, where each situation refers to two objects with all three attributes and relates to 113,064 unique utterances. Every situation is paired with at most 40 utterances.

2.1 Situation Representation

We define a formal representation of objects, their properties, and their relations within the environment using a predicate-argument structure referred to as a situation s. The s represents the full formal representation of the situation but we also use a partial one referred to as m. Each predicate p can have one or two arguments representing objects, denoted by t. For example, $p(t_1)$ represents a single-argument predicate, and $p(t_1, t_2)$ represents a two-argument predicate.

Single-argument predicates describe the properties of an object including $\{shape, size, color\}$ whereas two-argument predicates describe the relationships between objects including $\{left, above\}$. Shape predicates include hexagon ($he1$), star ($st1$), triangle ($tr1$), square ($sq1$), circle ($ci1$), and ellipse ($el1$). Colors include red ($re1$), blue ($bl1$), yellow ($ye1$), orange ($or1$), green ($gr1$), and purple ($pu1$). Sizes include small ($sm1$), medium ($me1$), and big ($bi1$). Relations include left ($le2$) and above ($ab2$). Although the relation predicates are limited to two relation types, these are sufficiently expressive for two additional relations: right and below. For utterances that include right and below, their formal situation will use the predicates $le2$ and $ab2$ respectively while the order of their arguments reflects the actual relationship.

Each *full situation* in the dataset includes all the properties of two objects and their relation. A partial situation representation can include a subset of the predicates. However, a valid partial situation must have at least one shape predicate to be able to refer to at least one object. We refer to the object types, attributes, and their relations as concepts to be learned. We gener-

ate all possible formal situations based on the possible combinations of the concepts given a set of templates. An example of a formal situation is:

```
sm1(t1),bl1(t1),ci1(t1),
ab2(t1,t2),
bi1(t2),re1(t2),sq1(t2)
```

where the following natural language expression, "the small blue circle above the big red square" is a valid explanation for it. In the generation process of formal situation representation, we consistently use the order of object/relation/object where each object is described as follows: size/color/shape.

2.2 Linguistic Utterances

A linguistic utterance, denoted as u, is a sequence of words used to describe a formal representation of situations mentioned in Section 2.1. All linguistic utterances in this data follow the grammar shown in Figure 2 with a vocabulary size of over 20 words. This grammar generate utterances like: "the small blue circle", "the star", and "the small yellow square above the medium red hexagon".

S →	<Object> | <Object> <UpDown> <Object> | <Object> to the <LeftRight> of <Object>
Object →	the <Size><Color><Shape>
Size →	big | medium | small | ϵ
Color →	red | blue | yellow | orange | green | purple | ϵ
Shape →	circle | triangle | square | hexagon | star | ellipse
UpDown →	above | below
LeftRight →	right | left

Figure 2: Grammar used for utterance generation.

2.3 Situation-Utterances Alignment

To connect all valid utterances from the grammar shown in Section 2.2 to a formal situation mentioned in Section 2.1, we define a meaning transducer T which receives an utterance u as input and generates a formal representation of the utterance m. An example of this transducer T is shown in Figure 3. Table 1 shows examples of linguistic utterances and their formal representations. For example, the utterance "the small circle" is mapped to shape and size predicates with the following predicates "$sm1(t_1), ci1(t_1)$", which is a valid formal representation generated by the transducer T. In addition of building the dataset for the experiments presented in this paper, the teacher model uses this transducer T twofold: (1) determine if the utterance follows the grammar and (2) generates the formal representation m from the learner utterance for feedback generation.

3 Language Acquisition Setup

The language acquisition framework presented here aims to evaluate the influence of corrective feedback in the development of a learner model's capacity to describe various situations as well as its association of each word to a corresponding predicate. The learner receives formal situation representations which it uses to produce a valid utterance that describes the situation. The teacher model employs various feedback strategies to choose an utterance from a set of valid utterances in order to address any possible errors that said utterance might have. Figure 4 shows an interaction between the learner and the teacher.

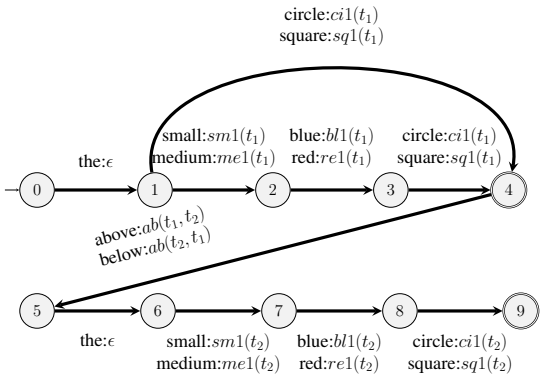


Figure 3: Meaning Transducer T. This is an example of the transducer which generates the formal representation of any utterance from the grammar.

3.1 Learner Model

The learner model is implemented as an Encoder-Decoder model with Gated Recurrent Units (GRU) (Cho et al., 2014; Sutskever et al., 2014) that receives a situation s which is a sequence of predicates as input and generates an utterance u which is a sequence of words describing the situation. We incorporate an attention module in the GRU architecture to improve learning the association between words and predicates (Bahdanau et al., 2016). The learner model uses cross-entropy (CE) loss between the learner utterance and the perceived utterances which are computed as follows:

$$l = -\sum_{t=1}^{T} \log p_w(y_t|u_{<t}, s) \quad (1)$$

where $y = y_1, \cdots, y_t$ is the ground truth utterance (which is selected based on the feedback strategy), u is the learner-generated utterance, and p_w, is the probability for generating the utterance given the situation parameterized with w.

3.2 Teacher Model

The teacher in this experiment is not a neural model rather is a predefined evaluator composed of an Evaluation and Feedback Generation modules. The former acquires the logical semantics of an utterance using the transducer defined in Section 2.3 to determine the validity of the utterance. The latter uses the logical semantics of the utterance and the situation to select a

Linguistic representation (u)	Formal representation (m)
"the small circle"	$sm1(t_1), ci1(t_1)$
"the small blue circle to the left of the big red square"	$sm1(t_1), bl1(t_1), ci1(t_1), le2(t_1, t_2), bi1(t_2), re1(t_2), sq1(t_2)$
"the red square to the right of the blue circle"	$bl1(t_1), ci1(t_1), le2(t_2, t_1), re1(t_2), sq1(t_2)$
"the circle to the left of the square"	$ci1(t_1), le2(t_1, t_2), sq1(t_2)$

Table 1: Examples of linguistic utterances and their formal representations related to the situation $s = sm1(t_1), bl1(t_1), ci1(t_1), le2(t_1, t_2), bi1(t_2), re1(t_2), sq1(t_2)$.

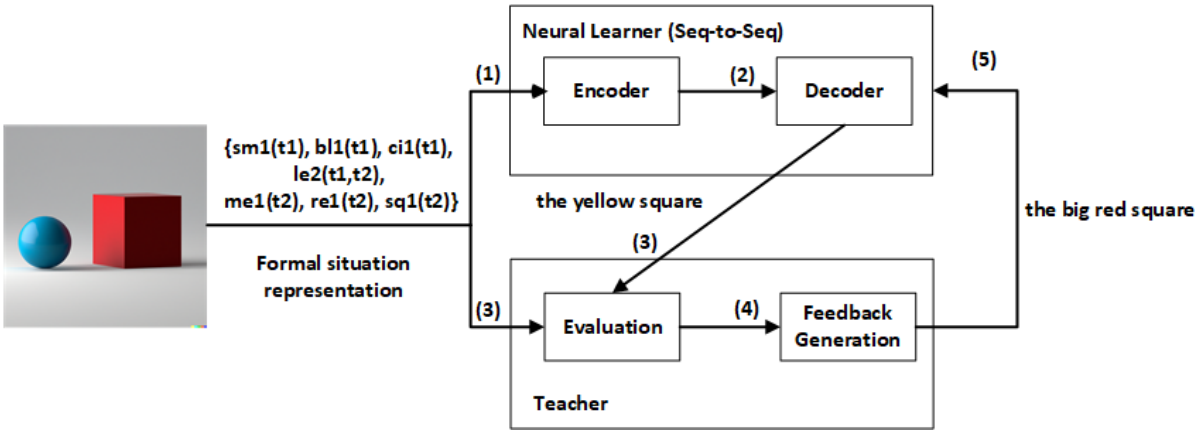


Figure 4: Interaction between learner and teacher models. (1) The situation's formal representation is passed to the learner's Encoder. The image shown here is for visualization purposes, each situation is written in formal representation. (2) The encoded situation is passed to the Decoder. (3) The generated utterance by the decoder, is passed to the Teacher's evaluation module along with the situation. (4) The feedback generation module receives the situation representation if any error is detected. (5) The feedback generation module generates an utterance following the feedback strategies mentioned in Section 3.3 so the learner uses it for training as a means of correction.

valid utterance corresponding to the situation.

Evaluation module. This module evaluates if an utterance is part of the set of valid utterances corresponding to the situation. It uses the transducer T to classify an utterance as "syntax error", "error in meaning", or "correct". An utterance is classified as "syntax error" when the transducer is unable to extract a corresponding meaning from the utterance, "error in meaning" when the meaning is extracted from the transducer correctly but the utterance is not part of the set of valid utterances for the current situation or "correct" otherwise.

Feedback Generation module. This module selects an utterance from the set of valid utterances from the situation to provide corrective feedback whenever an error is detected. This selection varies according to the feedback strategies mentioned below.

3.3 Feedback Strategies

For each situation, there are multiple semantically correct utterances that describe it. While the learner model generates an utterance, the teacher model faces the challenge of choosing an appropriate utterance to correct the learner, while the teacher is not aware of the learner's intention (Lee et al., 2010). For this paper, we propose three feedback strategies that the teacher uses to select an utterance from the set of valid utterances to provide feedback: full-length, random, and lexical distance. We analyze the effect these strategies have on the learner's performance. Each of these strategies shows different scenarios or outcomes of each interaction. Table 2 shows examples on what utterances the teacher provides according to each strategy.

Full-Length Feedback. This strategy presents an scenario where the teacher chooses an utterance that provides a complete description of the situation. A complete description contains two objects with all of its properties (i.e. size, color, and shape) and a relation between both objects. This allows a more efficient learning of the association between words and predicates present in the situation.

Random Feedback. This strategy presents a scenario where the teacher provides randomly selected corrections from a set of valid utterances corresponding to the situation. This strategy allows the learner model to be exposed to a wider distribution of possible utterances that describe the observed situation.

Lexical Distance Feedback. This strategy shows a targeted approach where the teacher identifies errors from the learner's utterance by measuring the

minimum edit distance or lexical distance between the learner utterance to some valid response, at a word-level, which it then chooses a correction that is grammatically different to the learner's utterance while preserving its intended meaning.

	Full-length FB	**Random FB**	**Lexical Distance FB**
Learner	the small red triangle to the right of the big yellow triangle	the big triangle	the small circle
Teacher	the small red triangle to the right of the big orange triangle	the triangle to the left of the small triangle	the small triangle

Table 2: Example of learner utterances and teacher utterances for the situation s = {bi1(t1), or1(t1) tr1(t1), le2(t1,t2), sm1(t2), re1(t2), tr1(t2)} *FB: Feedback.

Lexical Distance Feedback can be interpreted as a form of recast due to how it provides feedback while keeping the original meaning of the learner utterance. In terms of neural models, it helps to reduce larger penalty using cross-entropy loss. While Full-length feedback and Random Feedback can be interpreted as reformulations, these feedback strategies can provide utterances that could have a different meaning than the learner's utterance. With this, we explore whether providing these type of feedback affects the diversity of the learner's production of valid utterances. In other words, providing feedback that has a wider range of possible utterances like Random feedback allows the model to generate different valid utterances than target feedback like Lexical Distance Feedback which provides feedback with specific corrections.

3.4 Corrective Feedback Frequency

Another aspect that we are interested in exploring is how the frequency in which feedback is provided could affect the learner's utterance production and perception. To address this we propose two frequency approaches: (a) Corrective (CO) and (b) Non-Corrective (NC). Corrective feedback is the case which the teacher only provides a feedback utterance when the learner model generates an invalid utterance while Non-Corrective feedback is the case where the teacher always provides a feedback utterance regardless of the validity of the learner's utterance.

We want to observe whether the performance of the model could affirm the assumptions that providing feedback more frequently could lead to better performance. There are discussions that providing limited feedback on particular tasks is not sufficient to help an individual correct any observed errors, while providing too much feedback can overwhelm an individual and might lead them to make more errors. In our experiments, we want to observe whether providing corrective feedback only when the learner's utterance contained errors could lead to similar or better performance than providing non-corrective feedback. This could help generate training paradigms that focuses on addressing the frequency and quality of feedback which can reduce the amount of training data needed for language learning tasks.

4 Experimental Settings

All learner models were implemented using Pytorch (Paszke et al., 2019). The encoder and decoder GRUs have a hidden size of 300. The situation inputs are passed into a one-hot embedding layer forming a context vector representation. We used the Adam optimizer with a learning rate of 0.001. All learner models are trained with 18,000 situation-utterance set pairs and evaluated with 5,000 pairs. The learner's evaluation performance is recorded every 500 interactions.

4.1 Production setting

The goal of this setting is to observe how the feedback strategies and frequency of feedback affect the learner's utterance generation. We aim to address the following questions: (1) Can the learner provide a valid description of the situation? (2) Can the learner generate all possible utterances for any given situation? To address both of these questions, we propose two evaluation metrics: Semantic Accuracy and Completeness.

Semantic Accuracy measures the model's capacity to produce an utterance whose formal representation is accurate to the formal representation of the situation. In other words, we want the learner to generate utterances that preserve the meaning of the situation. We observe the model's utterance production development where it computes the loss with a single feedback utterance from the teacher model.

Completeness measures the rate of generated utterances that are present in the set of valid utterances. We want to see how many descriptions of a situation the learner model can learn during interactions with the teacher. The model uses a beam search approach to generate top-k utterances where $k = 40$ and the beam width to 22 (i.e. the vocabulary size). This technique has been applied to various tasks like story generation that require the generation of multiple sequences Fan et al. (2018); Holtzman et al. (2019).

4.2 Perception setting

For this setting, we evaluate the learner's capacity to choose utterances that correctly describe a given situation over other utterances that have at least one error in their description of the situation. Many psycholinguistic studies have used a two-alternative forced choice (2AFC) task to explore a child's preference for relevant objects for some linguistic stimuli (Gertner and Fisher, 2012; Bergelson and Swingley, 2012). This evaluation has been adapted to evaluate computational models

for semantic evaluation of language models as the one shown in Nikolaus and Fourtassi (2021b).

Using the situation-utterance set pairs discussed in Section 2.3, we create triplet pairs $x_i = (s_i, u_t, u_d)$, where s_i is the given situation, u_t is the target utterance which is randomly selected from the set of valid utterance for that situation, and u_d is a distractor utterance that is similar to the target utterance, except that it has one instance of the evaluated concept replaced with another instance of the same concept.

For this task, we compute the probability of the target utterance given the situation and the distractor utterance given the situation. If the probability for the target utterance is greater than the probability of the distractor utterance, we consider that the learner model has successfully understood the evaluated concept. We train the models with the same 18,000 situation-utterance set pairs as the ones used in the production setting. For each concept, we create a set of 5,000 triplets for evaluation. We record the learner's performance after every 500 interactions.

5 Results

Syntactic Errors. Our model does not show any significant amount of syntax errors after a few interactions. Figure 5 shows the number of utterances that were classified as "syntax error" by the teacher's evaluation module for all models during the first 200 interactions. After 30 interactions, we see that corrective random feedback, non-corrective random feedback and non-corrective lexical distance have all of its generated utterances during the evaluation classified as "syntax error". Later interactions, we see that syntax errors become non-existent because the model is able to generate utterances that follow the expected grammar. Each learner model is initialized with the same random seed to ensure the comparison on the effects of each feedback strategy are from the same starting point. This high syntax error after 30 interactions present a point where the models were starting to get utterances with repeated shape concepts at different position before learning that size and color concepts occur before shape concepts, thus adjusting to the grammar.

Semantic Accuracy. Figure 6 shows the learner's semantic accuracy for each feedback strategy and feedback frequency. It is observed that corrective random feedback and corrective lexical distance has higher semantic accuracy than all other models while non-corrective and corrective full-length feedback have the lowest semantic accuracy. There were cases where the learner model described the situation using one relation while the teacher provided a valid feedback with the opposite relation. This causes confusion to the model due to misaligned objects. For example, the learner generated the utterance "the big red star to the left of the small yellow circle" and the teacher provided "the small yellow circle to the right of the big red star". Corrective random feedback and lexical distance

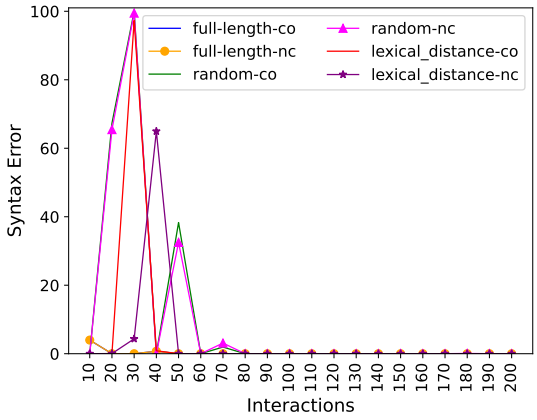


Figure 5: Syntax Errors for each feedback strategy with corrective and non-corrective feedback for the first 200 interactions. All models are evaluated after every 10 interactions with 5000 situation-utterance pairs.

shows oscillations between interactions. We observed that these feedback strategies provide utterances of different lengths more frequent than full-length feedback which causes the learner model to fail to describe some situations.

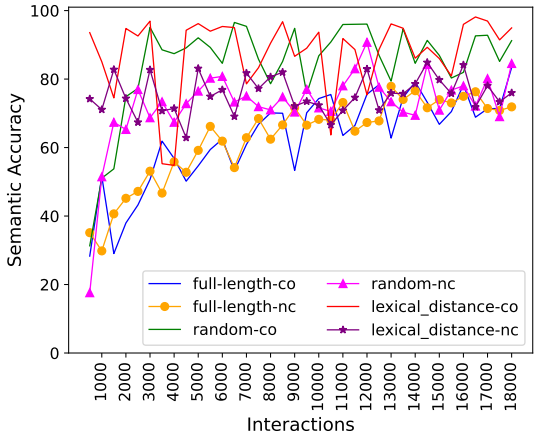


Figure 6: Semantic Accuracy evaluation for each feedback strategy with corrective and non-corrective feedback.

As shown in Table 3 most models reach an accuracy of sixty percent or more within the first 500 interactions for lexical distance, 1000 interactions for random feedback, and 3000 interactions for full-length feedback. Random feedback and lexical distance are the only strategies that reach ninety percent semantic accuracy. Corrective lexical distance feedback reaches ninety percent after 500 interactions.

The learner trained with lexical distance feedback strategy correctly generate valid utterances earlier than other strategies because these generated utterances only describe the shape of at least one object in the situation. On the other hand, full-length feedback expects the learner model to describe the size, color, and shape

of both objects which requires more interactions for the learner to successfully describe a situation following this strategy.

All feedback strategies with corrective frequency are able to reach each performance threshold with less interaction than non-corrective frequency. This is interesting because it shows that for this task, limiting the amount of corrective feedback helps the learner generate more valid utterances for each situation.

learner \ accuracy	60%	70%	80%	90%
full-length-co	3000-3500	7500-8000	17500-18000	-
full-length-nc	5000-5500	8500-9000	-	-
random-co	1500-2000	1500-2000	2500-3000	2500-3000
random-nc	1000-1500	2000-2500	5000-5500	11500-12000
lexical-distance-co	0-500	0-500	0-500	0-500
lexical-distance-nc	0-500	0-500	1000-1500	-

Table 3: Number of interactions where the learner model reached or exceeded certain semantic accuracy threshold percentage. Since each model is evaluated every 500 interactions, we show the interval range where the model reached the specified accuracy percentage.

Completeness. Figure 7 shows the learner completeness score at various interactions between the models. It shows that non-corrective feedback strategies have higher scores than corrective feedback strategies. Non-corrective random feedback has the highest completeness score at 55%, which means that the learner model trained with this strategy and frequency was able to generate around 20 different valid utterances. On the other hand, corrective full-length feedback only generated 3 valid utterances which is approximately 7%.

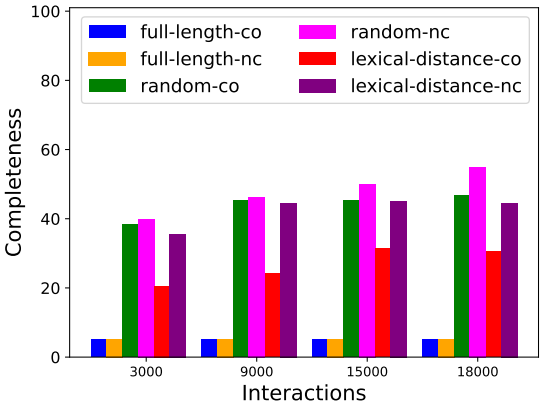


Figure 7: Completeness evaluation for each feedback strategy with corrective and non-corrective feedback.

When comparing the completeness score and the semantic accuracy, there is no indication that high semantic accuracy also means high completeness. Instead, we observed that full-length feedback had low semantic accuracy and low completeness score. Also, both random feedback and lexical distance feedback had high semantic accuracy and high completeness. A possible explanation for this lies on the diversity of feedback utterances the learner receives. Random feedback provides utterances of different lengths whereas full-length feedback only provides one or two utterances with the full description of the situation. We can also argue that Lexical distance feedback is in the middle in terms of performance because this strategy selected utterances with minimal edit-distance to the learner's utterance, thus limiting the range of possible utterances. We believe that exposing the learner to different utterances are helpful for the learner model to develop its set of utterances to describe each situation.

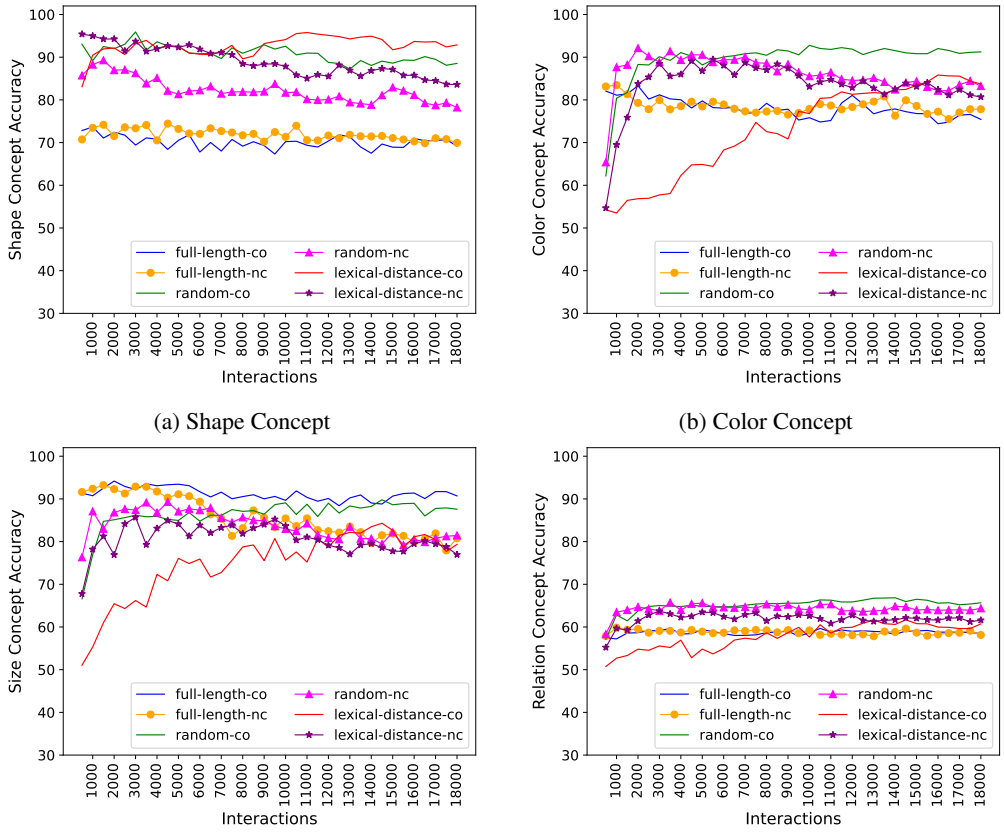
Perception Evaluation.

Figure 8 show the perception evaluation for every concept with each feedback strategy and feedback frequency. We observe that there is no significant difference between the corrective and non-corrective frequency of each feedback strategy. Corrective full-length feedback has the lowest shape accuracy and highest size accuracy. We observe that corrective lexical distance has low accuracy for color, size, and relation concept during early interactions but increases its accuracy to be on par with other feedback strategies.

All learner models are able to have concept accuracy above 50%, therefore we could argue that these models are able to learn these concepts. There are some cases where some models have difficulty learning certain concepts. One case is the relation concept where all models have low accuracy. For example, full-length feedback strategy, the learner has learned to generate an utterance like "the big red star to the left of the small orange circle" but the current concept evaluation test expects a target utterance like "the small orange circle to the right of the big red star". These utterances while both equally valid, the learner model might not understand the target utterance due to the positioning of the objects. For random and lexical feedback, the learner did not learn to generate utterances with relations.

6 Discussion

In the production setting, each feedback strategy presents forms in which a teacher could provide feedback to a learner as guided responses, from describing the full situation to rephrasing the utterance to correct incorrect or missing concepts. The learner models had a positive response to these guided types of responses which yielded higher semantic accuracy. However, we noticed that each model generated utterances of specific concept combinations. This is also supported by the completeness evaluation which shows the set of utterances the model is able to associate to each situation. Random feedback strategy generated utterances of different length for each situation. These results indicate that exposing the learner model to various feedback utterances across multiple situations can improve the model's semantic accuracy as well as completeness score.



(a) Shape Concept

(b) Color Concept

(c) Size Concept

(d) Relation Concept

Figure 8: Perception Task evaluation for all feedback strategies within the corrective and non-corrective setting.

In the perception setting, we measure a model's concept accuracy which indirectly shows how the model develops word-meaning associations when the model successfully selects a target utterance which has the correct use of a concept as well as correctly describing the current situation. Each feedback strategy present a constant concept accuracy throughout the interactions.

We analyzed aggregated attention maps during the model's production setting to measure the word-meaning mappings were developed as new interactions occurred. The maps for full-length feedback show that this strategy help the learner model have strong one-to-one association between concept words and their respective predicates. For random feedback and lexical distance feedback strategies, the models formed associations between each word to multiple predicates corresponding to different concepts. Further details and attention maps can be found in Appendix A

In addition to analyze attention maps, we observed the error rates of every concept for each feedback strategy. Overall, the error rates for each concept decreases as the learner interacts with the teacher. Some strategies like random and lexical distance feedback have higher error rates at earlier interactions, they still decrease during later interactions. Additional details are in Appendix B

Although the evaluation metrics in our experiments aim to explore the acquisition of a natural language by neural models, they are not sufficient and thus more complex metrics are required. For a fair comparison between neural language acquisition and human language acquisition, establishing cognitive plausible neural architectures. By cognitive plausible, we mean that neural models could emulate human-like processing (Beinborn and Hollenstein, 2023). We believe that this framework has potential to include additional metrics that can evaluate neural model learning and gain insights to language acquisition theories.

7 Conclusion

This work proposes a framework to explore teacher-learner interactions with corrective feedback within a controlled environment using formal representations of objects and their properties. We evaluate various feedback strategies and their influence on the learner model's utterance generation for a given situation and perception of different concepts like shape, color, size, and relation present in a situation. These results show that the learner models can generate a different subset of valid utterances to describe a situation according to the feedback strategy employed by the teacher. Some strategies like random and lexical distance were useful for the learner model to learn multiple utterances whereas full-length only allowed the learner model to learn 1-2 valid utterances. In terms of perception, some models present challenges for certain concepts due to unseen target utterances during concept evaluation that were not provided by the teacher during training interactions.

Given the reduced size of the vocabulary and the concepts, the use of GRU was selected to better highlight the main idea of creating a controlled framework that could be insightful, in terms of production and perception of a synthetic language, for the community to study the challenges of language acquisition and in-depth semantic evaluation of generated utterances. Further study and technical developments are needed to possibly incorporate newer neural models into this framework. In this paper, the teacher's feedback served as the ground truth labels used by the learner model. The learner model did not do any analysis or additional processing of the teacher's feedback which might limit the effectiveness of feedback and interactions For future work, we need to design procedures where the learner model can interpret the teacher's feedback and use said interpretation in the production process. One approach is to introduce mechanisms that allows the learner model to use the formal representation of the teacher's utterance to incorporate semantics into the loss function calculations.

8 Limitations

The framework has some limitations that need to be addressed. First, this framework uses synthetic data within a controlled environment. Natural language is very complex, which makes it a challenging task to create evaluation frameworks. Our data do not fully represent a natural language. Second, the proposed models to evaluate the data are basic in design. Our models rely on single-layer GRU with sequential formal representations of situations. We need to explore the use of this data on other models to have a wider panel of performance for better comparison and evaluation. Although our intent is to study child language acquisition, our work does not have any empirical analysis between the neural models and other studies involving actual children. Our goal is to explore additional metrics and evaluation settings in which this framework compares neural learner models with human performance.

References

Dana Angluin and Leonor Becerra-Bonache. 2017. A model of language learning with semantics and meaning-preserving corrections. *Artificial Intelligence*, 242:23 – 51.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Preprint*, arXiv:2106.08694.

L. Beinborn and N. Hollenstein. 2023. *Cognitive Plausibility in Natural Language Processing*. Syn-

thesis Lectures on Human Language Technologies. Springer International Publishing.

Tony Belpaeme and Anthony Morse. 2012. Word and category learning in a continuous semantic domain: Comparing cross-situational and interactive learning. *Advances in Complex Systems*, 15(03n04):1250031.

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.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Michelle M. Chouinard and Eve V. Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(3):637–669.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint*.

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

Jerome A Feldman, George Lakoff, Andreas Stolcke, and Susan Hollbach Weber. 1990. Miniature language acquisition: A touchstone for cognitive science. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pages 686–693. Citeseer.

Stefan L. Frank, Padraic Monaghan, and Chara Tsoukala. 2019. Neural network models of language acquisition and processing. In *Human Language: From Genes and Brains to Behavior*. The MIT Press.

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

Lila Gleitman. 1990. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55.

Sarah Hiller. 2016. Corrective feedback in first language acquisition.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Alexis Juven and Xavier Hinaut. 2020. Cross-situational learning with reservoir computing for language acquisition modelling. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

S.D. Krashen, S.D. Krashen, and T.D. Terrell. 1983. *The Natural Approach: Language Acquisition in the Classroom*. Language Teaching Methodology Series. Pergamon Press.

Patricia K. Kuhl, Feng-Ming Tsao, and Huei-Mei Liu. 2003. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, 100(15):9096–9101.

Sungjin Lee, Cheongjae Lee, Jonghoon Lee, Hyungjong Noh, and Gary Geunbae Lee. 2010. Intention-based corrective feedback generation using context-aware model. In *CSEDU*.

Michael H. Long. 1981. Input, interaction, and second-language acquisition. *Annals of the New York Academy of Sciences*, 379(1):259–278.

Roy Lyster and Leila Ranta. 1997. Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19(1):37–66.

Padraic Monaghan, Simón Ruiz, and Patrick Rebuschat. 2021. The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local and global effects on acquisition. *Second Language Research*, 37(2):261–289.

Mitja Nikolaus and Abdellah Fourtassi. 2021a. 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 and Abdellah Fourtassi. 2021b. Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 391–407, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Steven Pinker. 2013. *Learnability and Cognition: The Acquisition of Argument Structure (1989/2013)*, new edition edition. Cambridge, MA: MIT Press.

Eva Portelance and Masoud Jasbi. 2023. The roles of neural networks in language acquisition. *Retrieved from osf. io/preprints/psyarxiv/b6978*.

W. V. O. Quine. 1960. *Word & Object*. MIT Press.

J. Sachs, B. Bard, and M. Johnson. 1981. Language learning with restricted input: Case studies of two hearing children of deaf parents. *Applied Psycholinguistics*, 2:33–54.

Ted Schoneberger. 2010. Three myths from the language acquisition literature. *The Analysis of Verbal Behavior*, 26:107–131.

Kenny Smith, Andrew DM Smith, and Richard A Blythe. 2011. Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35(3):480–498.

Linda Smith and Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106:1558–68.

Luc Steels. 1995. A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332.

Sumarga Suanda, Nassali Mugwanya, and Laura Namy. 2014. Cross-situational statistical word learning in young children. *Journal of experimental child psychology*, 126.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Alex Warstadt and Samuel R. Bowman. 2024. What artificial neural networks can tell us about human language acquisition. *Preprint*, arXiv:2208.07998.

Chen Yu and Dana H Ballard. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)*, 1(1):57–80.

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

A Word Learning Evaluation

Specifically, it shows the attention maps for each feedback strategy and feedback frequency after 18,000 interactions.

The learner models trained with corrective full-length feedback has strong word-predicate associations for all concepts. For example, the model trained with corrective full-length feedback shows that the word "hexagon" has a strong association with the predicate "he1" while having little to no association to any other predicates. However, models trained with non-corrective full-length feedback show weaker associations between words and predicates in the sense that one word is associated to more than one predicate. For example, the word "big" is associated to the predicates "bi1"as well as other color predicates like "ye1", "gr1", and "re1". Also, these models showed strong associations for the relation concept words "above" and "below" to predicate "ab2" as defined in the grammar discussed in Section 2.2. Unfortunately, the model did not present any strong associations between the words "left" and "right" to the predicate "le2". This could occur because the model was not provided with sufficent feedback utterances that uses both of these relations.

Models trained with corrective random feedback, as shown in Figure 9b, have strong associations for the shape concepts. However, for color, size, and relations concepts, we see that the model has higher attention towards shape and relation predicates. For example, the word "red" has associations to color predicates like "st1", "ci1" and even relation predicates "ab2" and "le2". Possibly this model receives feedback utterances that might contain instances of color and size concepts, which are not required for an utterance to be valid, at different positions thus making it difficult to fully adjust the attention weights. In Figure 9e shows that non-corrective random feedback present the same issue.

Models trained with lexical distance feedback shows that most words had strong association to relation predicates. For example, the word "triangle" has high association to the predicates "le2" and "ab2". Also, we observe that words of shape instances have high association to its corresponding predicate. Similarly to random feedback, this strategy also has difficulty in properly update the attention weights for different concepts. The model associates the words for sizes like "small", "medium", and "big" to various shape predicates. The same is observed for color concept which also associate each color word to the shape predicates.

We can say that these attention maps are consistent with the behavior of each feedback strategy. Full-length feedback provide the full description of the situation which makes the attention between words and predicates simple to compute. Random and lexical distance feedback provide utterances with partial descriptions which can affect how certain words are aligned to the situation. For example, the utterance "the red circle" being aligned to the situation "bi1(x1) re1(x1) ci1(x1) le2(x1,x2) sm1(x2) or1(x2) tr1(x2)" might align the word "red" to the predicate "bi1" which occurs before the predicate "re1".

B Concept Error Rate Analysis.

We conducted an analysis of the frequency of concept errors during the production evaluation. Table 4 shows the error rates (i.e., the number of concept errors divided by the number of examples presented during evaluation) for each concept across different interaction periods. The relation concept had values lower than 1% therefore were not included in this table. These results indicate that the error rate trends depend on the feedback strategy employed as explained below.

Strategies with corrective feedback have a lower error rate than non-corrective feedback across all strategies for all concepts. Also, we observe that for shape and size concepts, full-length feedback has a decreasing error rate as it interacts with the teacher model. However this strategy has higher error rates for color concepts. We hypothesize that the model generate ut-

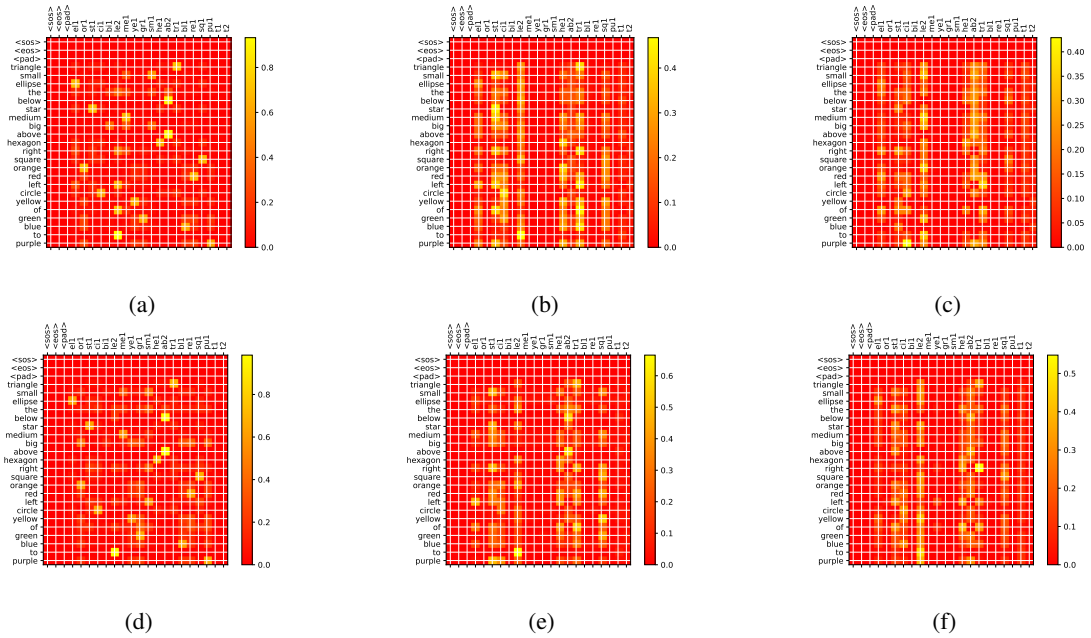


Figure 9: Aggregated attention maps for each feedback strategy and feedback frequency after all 18000 interactions. This values are the average attention weights for 5000 evaluation interactions. Attention maps for feedback strategies with corrective feedback, (a) Full-length feedback, (b) Random feedback, (c) Lexical Distance feedback. Attention maps for feedback strategies with non-corrective feedback, (d) Full-length feedback, (e) Random feedback, (f) Lexical Distance feedback.

terances with shape instances, followed by utterances with size and shape instances before learning color instances. Also, we have cases where the learner receives teacher feedback with the opposite relation from what the learner generated. This causes the learner to learn incorrect color instances for the intended object. Another thing to notice is that for random feedbaqck and lexical distance feedback strategies, the error rates tend to be lower than full-length feedback for early interactions while presenting higher error rates at later interactions. This occurs because both of these strategies are trained with partial descriptions of a situation, unlike full-length feedback, making them less prone to generate incorrect utterances.

For the relation concept, we observe that no errors were detected due to the model either not generating utterances that do not have any instances of relations or the relation were correctly generated. For example, a given situation could refer to a big star that is above a small yellow square as "the big red star above the small yellow square" rather than "the small yellow square below the big red star". These two sentences are both valid, however, their relation concepts refers to the objects in opposite perspectives.

This shows how these models are able to minimize the error rate in terms of utterance generation with the corresponding concepts as the interactions with the teacher increases. Although some models have a large error rate at the early interactions, we believe that these errors occur due to the utterances provided by the teacher model which could be valid to describe the given situation, but might not be fully aligned to what the learner model intended to describe. An example of this can be seen in the case of the color concept for the corrective full-length feedback strategy where the learner model attempts to describe "the big yellow circle" but the teacher model provides "the small blue star above the medium yellow circle" which by our CE-loss, it will train the learner to better learn to describe "the small blue star".

Shape Concept										
	500	**2500**	**4500**	**6500**	**8500**	**10500**	**12500**	**14500**	**16500**	**18000**
full-length-co	14.29	3.28	2.91	2.76	3.87	0.12	2.54	1.03	0.45	1.35
full-length-nc	11.87	3.11	1.01	3.18	1.93	1.64	1.15	1.83	3.14	1.87
random-co	15.00	9.37	1.65	7.51	1.21	0.76	1.88	2.93	0.75	0.34
random-nc	15.64	6.68	3.68	2.54	0.85	1.97	2.84	3.18	3.03	2.43
lexical-distance-co	15.22	8.15	12.11	8.27	14.20	2.62	6.87	6.30	8.87	6.82
lexical-distance-nc	11.06	9.55	12.70	9.95	5.39	6.73	4.11	8.08	9.49	7.60
Color Concept										
	500	**2500**	**4500**	**6500**	**8500**	**10500**	**12500**	**14500**	**16500**	**18000**
full-length-co	14.33	17.20	18.37	19.91	21.65	23.94	22.55	23.43	23.99	24.22
full-length-nc	12.83	18.97	21.36	21.14	22.81	22.62	23.23	23.36	22.57	23.28
random-co	10.07	3.89	5.88	13.01	5.59	4.19	6.97	9.83	2.24	7.69
random-nc	15.70	6.59	10.47	8.98	4.60	11.50	14.16	6.45	9.43	5.82
lexical-distance-co	7.53	3.44	0.17	1.92	8.80	11.80	6.47	12.78	5.65	1.48
lexical-distance-nc	13.40	6.62	5.94	5.16	13.72	9.12	11.78	5.21	6.45	10.46
Size Concept										
	500	**2500**	**4500**	**6500**	**8500**	**10500**	**12500**	**14500**	**16500**	**18000**
full-length-co	11.09	10.85	9.26	7.65	8.32	4.04	4.00	2.37	1.86	1.35
full-length-nc	12.07	8.66	6.42	5.23	2.41	4.10	3.97	2.27	3.29	2.81
random-co	8.92	14.16	18.66	6.07	20.05	20.97	18.03	13.51	20.53	17.48
random-nc	10.72	15.87	14.02	16.39	21.13	13.93	10.03	16.90	13.65	17.76
lexical-distance-co	3.18	6.87	4.18	7.66	3.86	13.97	12.10	6.02	11.02	16.01
lexical-distance-nc	4.89	9.89	10.29	10.92	8.36	10.13	9.68	11.94	9.54	7.50

Table 4: Error rates for each concept after training interactions. Measures the number of learner utterances classified as error in meaning due to the incorrect use of one of the concepts.