Grounding Novel Utterances in Visual Dialogue

Mert İnan and **Malihe Alikhani**
Computer Science Department,
School of Computing and Information,
University of Pittsburgh, Pittsburgh, USA
{mert.inan, malihe}@pitt.edu

Abstract

Interlocutors use sufficiently salient yet creative and dynamic meaning pairs to communicate and coordinate in dialogue (Lewis, 2008). In this work, we focus on novel utterances in visual dialogue. We survey different types of lexical innovations discussed in the cognitive science and computer science literature and study how and when the transformer-based language models fail to probe context and process novel referring expressions. We annotate around 300 utterances that include novel utterances from the Photobook dataset (Haber et al., 2019) and present a data-driven study of lexical innovation and micro language in task-oriented dialogue. We then propose an algorithm that ranks the importance of the local context history according to the content of novel utterances. Based on this ranking, we create a model that can process and ground these novel utterances in context. We conclude with a discussion on how lexical innovations may change across conversations and how interlocutors can converge on shorter referring expressions about 52% of the time over the course of the interaction.

1 Introduction

Communication is inherently creative. Interlocutors produce utterances that include novel expression–meaning pairs to successfully communicate (Clark and Clark, 1979). Listeners understand these *lexical innovations* and uncover the intended meaning effortlessly. We build on Armstrong (2016)'s argument and present empirical evidence that shows that semantic conventions that influence language production in dialogue are dynamically determined by coordination between the engaged listener and speaker. These local conventions, which Clark (1998) refers to as *micro-languages* are suited for the needs of subgroups and may not be utilized by other subgroups or even the same speaker or listener in future interactions.

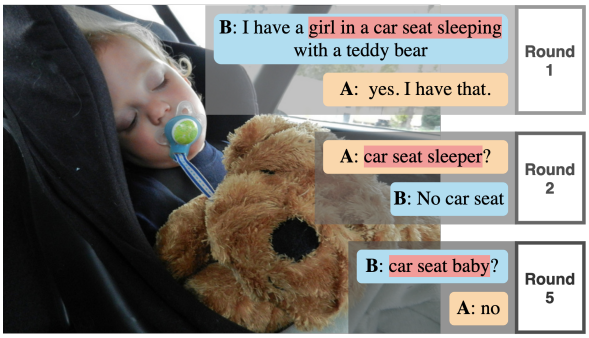

Figure 1: This is an example image from the Photobook Dataset, with its relevant dialogue history from different rounds of the game. The highlighted segments in the text correspond to the lexical innovations and their referents from the previous context. "car seat sleeper" is a novel utterance with noun-noun-noun compounding. Participants can converge to shorter lexical innovations over time. Converged lexical innovations can also change slightly, i.e, "car seat sleeper" to "car seat baby".

Krahmer and van Deemter (2012) discuss several ways in which the production of referring expressions can be creative and addressee–dependent i.e. the use of abbreviations and certain redundancies. In this work, we focus on visual dialogue and study ways that our cognitive capabilities and conventions might influence the production of lexical innovation. In particular, we evaluate the performance of language understanding models when conversations involve novel utterances in a visual dialogue task.

Resolving novel referring expressions in visual dialogue requires understanding the images and the context of the conversation. Agarwal et al. (2020) show that transformer models fail to understand and use the context in visual dialogue. They fail to generalize well in the presence of lexical innovations. We present a case study using the Photobook dataset (Haber et al., 2019) and examine the different types of lexical innovations that the dataset presents. Figure 1 presents an example conversation with lexical innovations and ways that the in-

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

terlocutors coordinate to understand each other and achieve a common ground. In what follows we (1) survey different definitions and classes of lexical innovations in the cognitive science and computer science literature; (2) present a data-driven analysis of lexical innovations in the Photobook dataset; (3) propose a novelty coefficient calculation algorithm which uses Part-of-Speech tagging to rank the novelty of utterances in a sentence; (4) develop a multimodal language understanding model that can detect and quantify novelties in the utterance, which then ranks local conversational context for better grounding of novelties; and, (5) show that lexical innovations are dynamic structures that are influenced by local conventions and may or may not be used in different dialogue rounds even by the same speakers.

2 Background and Definitions

Lexical Innovation Lexical innovation is different from conversational implicature (Bach, 1994), it is also not necessarily a deep metaphor. A deep metaphor is one, as defined in Armstrong (2016) and Lepore et al. (2010), where the meaning of the metaphor is linked to conventions that cannot be localized to that specific sentence. Furthermore, novel uses of utterances (most probably denominal verbs) that disrespect the semantic conventions are also not lexical innovations, such as "she foodified the ingredients", or "paper-outted the digital copy".

"Lexical innovation compositionally interacts with the other expressions in the sentences they occur", such as conditionals, negations and quantifiers. Armstrong (2016) argues that lexical innovation can happen in two ways:

- *Lexical expansion*: $L \rightarrow L'$ where L' is a lexicon with a new lexical expression that is not present in L.
- *Lexical shift*: $L \rightarrow L'$ where L' is a lexicon with a new lexical expression that is not present in L. Meaning of some expression in lexicon L' is different from that of in L.

In the fields of developmental psychology, and cognitive psychology, there are similar ways of defining strategies of lexical innovations. Clark (1980) and Bryant (2014) give the following strategies:

- Compounding: Two or more bases are combined and form a single word, for example, *bubble-hair* (a person with curly hair). Examples from Table 1 are *suit guy*, *truck thing*.
- Conversion: A word is adapted to a new word class without any changes. Examples from the Photobook dataset are *paving truck*, where paving is used as a noun instead of a verb or adjective.
- Affixation: A prefix or suffix is added to modify a base semantically or grammatically, for example, *sworder* (swordsman), *unfilled* (empty). Examples from the Photobook dataset are *pinkish*.
- Compounding with affixation: This is a subcategory of compounding and a combination of affixation. Examples from Table 1 are *the stripey cake*.
- Onomatopoeia: Words that sound like an action or an object. Examples from the Photobook dataset are *chuck*, *fluff*, *clip*, *pat*, *puff*.
- Child-talk: Such as that found in children's books (e.g., "Do you know what shlom is?"). Annotated section of the Photobook dataset does not contain child-talk due to its domain.

People are capable of producing novel utterances dynamically during a conversation. Yet, language understanding models fail to understand them (Testoni et al., 2022).

As an initial exploration, we focus on compounding and conversion—which are the major two categories of innovation that are automatically detectable—by using a detection algorithm that we propose. Using these strategies, we quantify the novel segments in grounded collaborative dialogues in a multimodal setting. We combine our understanding of lexical innovation from the already-present literature of linguistics, philosophy and cognitive psychology, and explore the Photobook dataset with the tools and understanding from these fields.

Collaborative Reference Grounding vague and ambiguous utterances have been addressed before in the context of grounding color terms. McMahan and Stone (2015); Monroe et al. (2017); Winn and Muresan (2018), and Fried et al. (2021) have all explored modelling color perception and comparative color descriptions using Bayesian models and reinforcement learning (Khalid et al., 2020a,b),

		
Utterance Chain	**Round 3:** **A:** my last one is the atari person with socks **Round 4:** **A:** okay, atari guy again **Round 5:** **A:** atari with socks	**Round 3:** **A:** I have the two men cutting a cake with red and white stripes **Round 4:** **B:** I have the two men with the stripey cake
Qualitative Observation	Across multiple rounds, utterance length becomes shorter, as the previous dialogue history context gets longer for both agents, the use of compounding to create novel segments increases. This example signifies looking up from previous history and increasing the attention to previous round utterances.	Across multiple rounds, different users can refer to the same object and only one of them may contain lexical innovation. While probing for context, both agents' previous rounds and previous turns should be used to find the necessary context for the specific novel segment.
		
Utterance Chain	**Round 1:** **B:** white guy with a orange vase looking at a truck thing **Round 3:** **B:** man in orange vase looking at a truck	**Round 3:** **B:** yes, I have the suit guy, on the bench again too **Round 4:** **B:** do you have the sloucher on the bench on his phone? **Round 5:** **A:** sloucher dude?
Qualitative Observation	Grammatical errors are not necessarily considered as lexical innovation. The lexical innovation detection module needs to be robust for these cases. Here, "vase" is not lexical innovation, but "truck thing" is.	The lexical innovation segment in the last round exists in the previous rounds as a part of the whole utterance. Hence, the segment needs to be extracted and then a coefficient needs to be calculated for the whole utterance to probe the previous context.

Table 1: This figure illustrates different dialogue examples based on the images above them. It also gives qualitative observations on how the lexical innovation plays a role in understanding certain segments of the utterances by the agents. Underlined portions denote the novel segments and their previous references. Most of them are noun-noun compounding lexical innovations.

whereas in our work we study Transformer-based models. While these papers are applied specifically to color terms and mainly work on resolving ambiguities, we are looking at grounding novel combinations of nouns that are not necessarily vague.

Resolving ambiguous novel utterances have also been studied in robotics and situated dialogue. It is still an open investigation area which has been mentioned in the recent survey for spoken interactions with robots by Marge et al. (2022). Liu et al. (2013) study novel referring expressions, where a graph mapping between a robot's visual context and the dialogue utterances is established for novel objects in the environment. In this line of research, a resolution of ambiguity of "novel" utterances have been addressed using cognitive processes. Our work is also inspired by the categories that cognitive scientists have proposed but we mainly focus on dynamically-formed novel utterances or micro language in visual dialogue.

Different corpora exist for the problem of visual collaborative reference: task-oriented visual dialogue such as *VisDial* (Das et al., 2017), *TalkTheWalk* where participants describe locations as they are walking, (de Vries et al., 2018), *MeetUp!* which is about dialogues that contain referring to locations and objects, (Ilinykh et al., 2019), *CoDraw* which has referring to objects and figures in drawings, (Kim et al., 2019), *Photobook* that has rich referring expressions to objects in a synchronous image matching game (Haber et al., 2019), *TEACh* where a commander directs a robot to complete tasks (Padmakumar et al., 2021), and SIMMC 2.0 (Kottur et al., 2021) where an agent resolves ambiguities when a human refers to objects in a shopping setting. While in all of these works there is

an exploration of resolving ambiguous referring utterances, none of the baseline models in these works address lexical innovations, and do not generalize well to out-of-domain corpora (Kim et al., 2020). Grounding and the problem of collaborative reference in dialogue is analyzed even more in the surveys by Schlangen (2019); Agarwal et al. (2020). Overall most models focus on a plethora of tasks and specific domains, but we are focusing on grounding creative utterances when people are referring to objects dynamically.

3 Data creation and annotation

In this work, we use the Photobook Task and its related datasets[1], which are components of a dialogue-based image-identification game (Haber et al., 2019).

In the original Photobook task, two participants are each shown 6 images selected from the MS COCO Dataset (Lin et al., 2014) on a randomized grid with some shared images. The primary task of the game is for each participant to select if any of the highlighted images is common or different by communicating with each other over a dialogue interface. The task is symmetric, as both participants can ask questions and provide answers. When the participants finalize a selection about the common or different images, then one round of the game ends, and another round begins with a newly randomized set of images. This new set may contain some of the same images from the previous rounds providing a history for participants to refer back to across rounds. A single game consists of five rounds, each of which contains three highlighted target images to label as common or different. This multi-round structure of the game allows an analysis of novel expressions that are getting created across different rounds by same or different participants, letting us observe the dynamics of lexical innovation. See Table 1 for a few examples.

Full dialogues of the Photobook dataset contain a total of 2,506 human–human conversations, and a total of 164,615 utterances. Because it is more straightforward to find novel utterances in the reference chains, we used that instead of the full dialogues. These chains are extracted from the full dialogues and for each MS COCO image in the game there is a chain. They are composed of multiple utterances taken from different rounds and different games referring to the same image. Each of these utterances contain a description about their corresponding image target from the dialogues. This Photobook utterance-based reference chain dataset is accessed through this link[2]. The total number of utterance chains is 16,525, which contain a total of 41,340 referring utterances. These are split into train, validation and test sets originally in the data with 11540, 2503, and 2482 utterances in each split, respectively.

As shown in Table 1 we observe various novel referring expressions such as "atari guy", and "the stripey cake" in the utterance chains. We formalize different classes of these type of novel referring expressions in Section 2, then annotate a portion of the utterance chains by identifying novel utterance segments and their classes. Then we use the Part-of-Speech tag patterns to detect these lexical innovations in Section 3.1.

3.1 Lexical Innovation Statistics in the Photobook Dataset

We observe that lexical innovation happens following semi-structured patterns of part-of-speech for the compounding and non-structured patterns for conversion classes. These patterns are as follows: for compounding, multiple `NOUN` classes are used consecutively; for conversion, an `ADJ` class or a `VERB` class is used in front of multiple consecutive `NOUN` classes. This is an empirical observation made on the available data, and it is assumed that these patterns are generalizable across datasets from different domains.

Counts for lexical innovation that we have identified in the Photobook dataset are presented in Table 2. This table shows multiple characteristics of the Photobook dataset in terms of lexical innovation. It shows that the most common way of creating novel words is by compounding. All the noun compoundings are the most common among all lexical innovation types. `ADJ-NOUN` and `VERB-NOUN` compoundings are assumed to be corresponding to the *conversion* type of lexical innovation.

We annotate a small subset of the training data (277 samples) with the lexical innovation types, by two human experts. In this data, we identify the presence of the segment inside the utterance that the lexical innovation is corresponding to. To calculate the Cohen's κ inter-rater agreement, we

[1] https://dmg-photobook.github.io/

[2] https://github.com/dmg-photobook/ref-gen-photobook/blob/main/dataset/v2.zip?raw=true

Lexical Innovation Type	Train	Test	Annotated
2-noun	4708	950	78
3-noun	1072	200	32
4-noun	202	37	2
5-noun	47	9	0
6-noun	99	29	0
adj-noun-noun	2981	662	46
verb-noun-noun	4471	121	14
onomatopoeia	1849	-	62
child-talk	-	-	0
affixation	-	-	0
Total count	34903	7450	277

Table 2: Numbers of different compounding types from the Photobook utterance chains. Annotated set is from the training set of the corpus. Dashes mean that POS-tag rules were not found to detect lexical innovation automatically in the data.

select 30 utterances randomly and assign them to two annotators. The Kappa coefficient is κ = 0.76 which indicates a substantial agreement (Viera et al., 2005).

After this annotation is complete, we run the automatic POS-tagging on this small subset. Here we observe that majority of the lexical innovations exist within the noun-noun compounding type. We observe that the distribution of the 2-noun, 3-noun, 4-noun, 5-noun and 6-noun compoundings follow a similar pattern for the train, test and the annotated subset. This shows that our POS-tagging strategy is a fast and feasible approximation of detecting lexical innovations similar to human annotations. We detect the onomatopoeia using a dictionary extracted from the Oxford English Dictionary by (Sugahara, 2011). After identifying these POS statistics, we try to find a way to quantify the novelty of these specific lexical innovation segments.

4 Model

Here we describe a listener model for collaborative reference grounding in the presence of novel utterances (see Figure 2). The inputs to our model are six images, one current utterance, and a history of reference-chain utterances referring to each of the six images, while the output is a single image chosen out of the six images. We measure our task success using accuracy and mean reciprocal rank (MRR) measures for image retrieval. We also present an algorithm for lexical innovation detection and coefficient calculation. Our code is publicly available[3].

Our model contains a modified listener module of the Reference Resolution Model as proposed by (Takmaz et al., 2020) [4]. In the original model, when the hypothesis utterance, u_t is received by the listener, BERT embeddings, $BERT(u_t)$, are extracted for each utterance using uncased base BERT (Devlin et al., 2019; Wolf et al., 2020) and they are concatenated with ResNet-152 embeddings, $RESNET(I_i)$ (He et al., 2016) of each of the images for multimodal representation. In our model, when the hypothesis utterance is received by the listener, we first identify whether there is a probable lexical innovation in the utterance. If there is no lexical innovation, then we run the original listener model. If there is, then we use a separate mechanism to rank previous rounds' utterances and increase the visibility of utterances that have less novel segments to the model.

When a lexical innovation is detected in the utterance, u_t, that refers to an image, i, then the model first fetches all the utterances that refer to i from the same game but previous rounds, which can be represented by $[u_{t-1}, u_{t-2}, ..., u_{t-k}]$, where t represents the current round number, and k is the number of maximum possible history of rounds for that specific image, i. Then we use our lexical novelty coefficient calculation algorithm to measure how novel each utterance in the history, u_{t-m}, is, where m is an arbitrary number less than k.

The novelty coefficient calculation is given in Algorithm 1. Given an utterance, we first run a POS-tagger on each word of the utterance, then find the segment of the utterance where it has POS tags corresponding to the segment of lexical innovation. We query Google n-gram Book database (Breder Birkenes et al., 2015; Lin et al., 2012) for that segment concatenated with its POS-tags (i.e. `umbrella_NOUN cat_NOUN lady_NOUN`) with insensitive case-matching and zero smoothing, resulting in a ratio, in the range of $(0, 1)$. If there was no lexical innovation segment found then we assign it a value of 1. We then multiply the ratio with the number of total entries in the Google Books n-gram database (around 10^{14} entries) to get an estimate count of the occur-

[3] https://github.com/Merterm/lexical-innovation

[4] model code is retrieved through: https://github.com/dmg-photobook/ref-gen-photobook/tree/main/models/listener

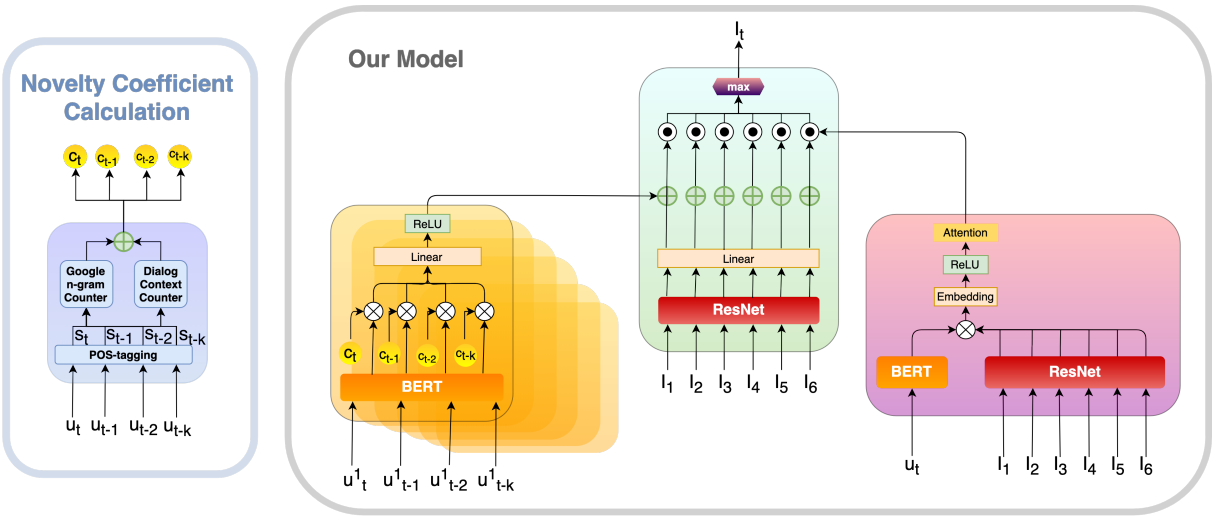

Figure 2: This is a simplified depiction of our multimodal language understanding model. Novelty weighing mechanism changes the weight of the history of utterances based on their lexical novelty during the dot product multiplication with each of the images. On the left, we have a lexical innovation coefficient calculator, which takes the dialogue history, $u^i_{t...t-k}$, and the current utterance, u_t, as input. It outputs coefficients, $c_{t...t-k}$, that are used in weighing the multimodal context in the model. In our model, there are 6 history modules per image in orange color, one utterance module in pink color, and one image picker module in green color. Inputs to the model are the utterance, u_t, history of utterances, $u^i_{t...t-k}$, for each image i, and six images $I_{1...6}$. The output is the chosen image, I_t, out of the 6 images. Intuitively, our lexical innovation module probes for the least innovative context and increases the weight of it in companion with the novel utterance.

rences of the segment in the English book corpora. This represents our game-independent external context coefficient with a range of $(0, 10^{14})$. Then we calculate the number of total occurrences of the segment in the given game and round, which represents our game-only local context coefficient. We finally calculate the novelty coefficient by adding both external and local context. The higher the coefficient, the less novel the utterance is.

Algorithm 1: Lexical Novelty Coefficient Calculation

Data: u_i, belonging to game g_i and round r_i
LI-RULES $\leftarrow$ {`2-NOUN, 3-NOUN, 4-NOUN, 5-NOUN, 6-NOUN, ADJ-NOUN-NOUN, VERB-NOUN-NOUN`}
for *each word,* w_i*, in* u_i **do**
  $tagged[w_i] \leftarrow$ pos_tagger(w_i);
end
if ***any*** *LI-RULES* ***in*** *tagged[i...i+n]* **then**
  $segment \leftarrow w_{i...i+n}$
end
$r_g \leftarrow 1$;
$q \leftarrow$ query_ngram($tagged$, $segment$);
if $q > 0$ **then**
  $r_g \leftarrow q$
end
$c_E \leftarrow r_g \times 10^{14}$;
$c_L \leftarrow$ count($segment$, g_i, r_i);
return $c \leftarrow c_E + c_L$;

In the original model, to pick one image out of 6 candidate images, a dot product is taken between the multimodal representation of u_t with the multimodal representation of each image $I_{1...6}$. The multimodal representation of u_t is the concatenation of ResNet features of $I_1...I_6$ with $BERT(u_t)$. The multimodal representation of each image, $I_i, i = 1..6$, is the summation of average of the history of the utterances, $BERT(u^i_{t-1}) + ... + BERT(u^i_{t-k})/k$, with $RESNET(I_i)$.

In our model, in order to pick one image out of 6 candidate images, we take a dot product between the multimodal representation of u_t with the multimodal representation of each image $I_{1...6}$, weighed by the lexical novelty coefficient. The multimodal representation of u_t is again the concatenation of $RESNET(I_1)...RESNET(I_6)$ with $BERT(u_t)$. However, in our case, the novelty-weighed multimodal representation of each image is the summation of the novelty-weighed history of the utterances, $c^i_{t-1} * BERT(u^i_{t-1}) + ... + c^i_{t-1} * BERT(u^i_{t-k})$, with $RESNET(I_i)$, where c^i_{t-1} represents the lexical novelty coefficient corresponding to u^i_{t-1}. Our model is depicted in Figure 2.

The main motivation for coefficient multiplication is to weigh the representations of different utterances from different rounds of the game. For

instance, if u_3^2–which is the utterance corresponding to round 3 of image 2–is "I have the two men cutting a cake with red and white stripe", and u_4^2 is "I have the two men with the stripey cake", then $c_3^2 > c_4^2$ as the u_3^2 does not contain any lexical novelty. We are giving highest weight to the least novel utterance because it is assumed that the least novel utterance representation is already grounded by the model compared to the most novel, which can guide the dot product towards grounding the novel utterance, as well.

	Full Photobook				**Only Novel**	
	Train		Test		Test	
	ACC	MRR	ACC	MRR	ACC	MRR
ReRef	95.2	97.3	85.3	91.2	82.5	89.5
Ours	97.6	98.7	85.4	92.1	**85.3**	**91.1**

Table 3: This table shows the performance of the Re-Ref model and our model on the train and test sets of the Full Photobook Corpus and Lexical Innovation-Only dataset. ACC corresponds to Accuracy and MRR corresponds to mean reciprocal rank. Re-Ref model performs worse on the lexical innovation extracted subset of the data than the full data. Our model improves on this giving more weight to less innovative utterances from the history of the conversation.

5 Results

We show that our model that is aware of lexical innovation improves on the accuracy and mean reciprocal rank (MRR) in the image retrieval task of choosing the target image from 6 candidate images (see Table 3).

Here we compare our model to the Re-Ref model introduced by (Takmaz et al., 2020). They show that their model performs with 85.32% and 91.20% accuracy in the test set of the full Photobook corpus. But we identify that their model's performance is slightly worse for the specific subset of lexical-innovation-only samples. As explained in Section 3, we select the samples using the automatic POS-tagger algorithm which contain segments that have lexical innovations in them according to our definition in Section 2. We show that ReRef model has an accuracy of 82.46% and an MRR of 89.49%, which are 3% and 2% less than the full dataset results, respectively.

Our listener model improves on the training data with around 2% in accuracy and MRR compared to the ReRef Baseline. More so, our model is able to bring up the test results for the lexical-innovation-only subset of the corpus to the full corpus performance levels. It improves the results by 2.8% for the accuracy, and 1.6% for the MRR compared to the ReRef baseline. In order to further investigate the performance of our model and investigate the dynamics of lexical innovation, we present qualitative and quantitative analyses in the following subsections.

5.1 Qualitative Error Analysis

In this section, the authors of the paper qualitatively observe the outputs of the novelty calculation. We see that the majority of the time, lexical innovation coefficient calculation successfully detects the novel utterances even in complex cases of 6-noun compounding. It is also able to detect non-novel utterances majority of the time as well.

We give more specific analysis of different types of qualitative phenomena we observe in Table 4. We can also see where the coefficient calculator does not perform as expected. For instance, one can observe that even though the sentences contain novel segments, the POS-tagging may select the non-novel segment such as "black bowl" instead of "orangy food", resulting in a false segmentation but correct coefficient calculation. This is still valuable for the listener model because the coefficient corresponds to all of the utterance instead of just the segment.

In certain cases, not novel segments can falsely get low coefficients (i.e. very novel), such as "hot dogs". This may be because Google n-gram database does not contain daily dialogues, and words that are not novel in daily communication may be absent on a book dataset, giving it a high coefficient even though it is not externally novel.

5.2 Do Lexical Innovations Change Across Games and Participants?

Lexical innovations can dynamically change during different rounds in a single game, during different games, and across different participants. Based on these three levels we ask three questions: how do novel words get modified across different rounds, how do they change across different games without considering rounds, and how do they change across participants regardless of the games? According to Armstrong (2016), lexical innovations exist dynamically, hence it can be hypothesized that after the game is over or even across different rounds, lexical innovation segments may get altered. To test this

	Utterance	Segment	Novelty Coefficient
Novel & Mis-segmented	do you have black bowl with orangy food, bowl with white rice, 3 part tray with food?	black bowl	1
	do you have a salad in a white bowl; salad looks like twigs with a red thing at the top...	white bowl	1
Novel & Correctly Segmented	green leafy salad with maybe red or orange item at top?	green leafy salad	1
	halloween cat?	halloween cat	1
	yes, pink rice, cat, tree, moon.	pink rice	1
	i have the red orange one	red orange	1
Not Novel with Low Coefficient	do you have a photo of fries and 3 hot dogs?	hot dogs	1
	black cat?	black cat	1
Novel with Medium Coefficient	do you have a dish on a square plate that has broccoli and white fluffy stuff ?	white fluffy stuff	4.03×10^6
	do you have broccoli with the white stuff again	white stuff	6.61×10^8
Not Novel with High Coefficient	salad with glass of grape juice or wine	-	10^{14}
	i have a picture with fries and three subs	-	10^{14}
Both Novel & Non-Novel with Different Scores	bowl of red vegetable next to loaf of bread on kitchen table?	-	10^{14}
	bowl of red veg next to loaf of bread	red veg	1
	a lunch box with 4 different colored comparents	colored comparents	1
	i have the lunchbox with the four compartments	-	10^{14}
Typo	largew hite square plate, with broccoli and rice etc	largew hite square	1

Table 4: This table shows different utterance examples and how the lexical innovation calculator module scores them for error analysis purposes. There are several classes of scoring and utterance pairs. First rows show differences in segmentation performance and how it affects the scoring. Next rows show how the novelty affects the score and finally an example with a typo is given. Here, higher score means less novel, as the novelty coefficient corresponds to a count of the word in the Google n-gram database and the previous dialogue context. Minimum score is 1, and the maximum score is 10^{14}.

hypothesis, we both qualitatively and quantitatively analyze the data. We list the lexical innovation segments that are found in our annotated data, then we cross-check the exact segment in our full dataset of utterances.

We find that lexical innovations re-occur in other games 22.2% of the time (267 different game re-occurrences out of 1203 lexical innovation re-occurences in the annotated dataset). This shows that same lexical innovation can be used multiple times across rounds and games. In Table 5, we present lexical innovation segment examples to observe their dynamic behavior across rounds and games. For instance, "white lap" re-occurs in different rounds of game number 744, 10.6% times out of all its re-occurrences. This shows that across different rounds, participants come back to the exact same lexical innovation segment. This is statistically significant with p = 0.0008 and t = 9.1259. We measure the significance using one sample t-test between the hypothetical uniform distribution mean of 1.52 of and the actual distributions across the games.

Lexical innovations from the annotated set re-occur 13.9% (167 same photo re-occurrences out of all lexical innovations re-occurrences) times when the picture is the same. Hence, different participants looking at the same picture can come up with the same lexical innovation even across different games. As an example, if we look at the same-photo re-occurrence probability of "choc cake", we see that 50% of its re-occurrence happens in games with the same photo, but with different participants.

On the contrary, participants can also converge to different lexical innovations when the game changes or after different rounds. For instance, in game 1140, participants can converge to "wii lap showing feet guy", then converge to "point of view wii remote" in another game. In another game, in round 3 participants converge to "feet up gaming" while in round 5 they re-converge to "close up wii remote guy". This shows that the durability of

Lexical Innovation	Game ID	In-Game Re-Occurrence Probability	Same-Photo Re-Occurrence Probability
choc cake	702	0.333‡	0.167‡
	635	0.111‡	0.278‡
	1900	0.167‡	0.500‡
	1903	0.111‡	
white lap	2433	0.091†	0.409‡
	1716	0.076†	0.136‡
	1346	0.061†	0.061‡
	2484	0.091†	0.242‡
	744	0.106†	0.061‡
salvation army truck	1502	0.081†	0.973
	1520	0.054†	
	1799	0.081†	
	2092	0.081†	
weird looking 5 wheeled black bike	1339	1.000	1.000
wii lap showing feet guy	1140	1.000	1.000

Table 5: This table shows the number of re-occurrences of some lexical innovation examples that were identified during annotation. In-game re-occurrence probability is the count of lexical innovation in the game with the given ID, divided by the number of total re-occurrences in all the annotated data. Same-photo re-occurrence probability is the count of the lexical innovation segment referring to the same photo divided by the number of total re-occurrences in all the annotated data. (†: statistically significant results with the power of $p \leq 0.001$, ‡: significant results with the power of $p \leq 0.1$)

novel utterances is dynamic, as some lexical innovations are easy for people to converge to and stay attached to even across rounds and games while some lexical innovations can dynamically vanish once the image or game is gone.

5.3 Do Participants Converge to Shorter Lexical Innovations?

People tend to converge to lexical innovations over the course of the dialogue in two different ways: either long and complex compoundings, or short and simple compoundings. Here we explore how these are distributed in our data. We observe that complex or lengthy lexical innovations that are 4 to 6-noun compounding do not re-occur in the data at all. These type of complex lexical innovations happen 46.9% of the time (130 out of 277 utterances) in the annotated dataset. We explain this phenomenon further with examples from Table 5. "weird looking 5 wheeled black bike" has an in-game re-occurrence probability of 1, which means that it only occurs in game 1339 once and never again in the data. This is because it is long and specific. For "wii lap showing feet guy", the participants converge to that lexical innovation in round 5 of the game, but it never exists in any other game. This shows the ephemeral nature of long and specific lexical innovations.

We observe another phenomenon in which participants converge to simpler and shorter lexical innovations as they continue to future rounds. In the annotated dialogues with lexical innovation, 51.8% (28 games out of 54) of the games converge from more than 5-token description of the object to 2 or 3-noun compounding lexical innovation after multiple rounds. This shows that participants converge to shorter lexical innovations as an establishment of common ground.

6 Discussion & Conclusion

We introduce a language understanding model that is able to probe both previous dialogue context and the external context for grounding novel utterances. The proposed model performs better particulary on the subset of the data that includes lexical innovations. Due to the nature of the task, users tend to come up with similar "novel" segments. Hence in the end, task-specific models which just memorize the vocabulary can perform just as well as a lexical-innovation-aware model. Also, as is shown in Table 2, it is difficult to find POS rules for lexical innovations, and some lexical innovation types such as child-talk do not exist in our chosen multimodal dataset, which requires further data exploration. Exploring other multimodal dialogue corpora is left for future work.

7 Acknowledgements

We thank Katherine Atwell, Anthony Sicilia, Ece Takmaz, Vishakh Padmakumar, Joshua-Christian Wyatt and Sabit Hassan and the anonymous reviewers for the helpful comments and discussions.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Josh Armstrong. 2016. The problem of lexical innovation. *Linguistics and Philosophy*.

Kent Bach. 1994. Conversational impliciture. *Mind & Language*, 9(2):124–162.

Magnus Breder Birkenes, Lars G. Johnsen, Arne Martinus Lindstad, and Johanne Ostad. 2015. From digital library to n-grams: NB n-gram. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 293–295, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Judith Becker Bryant. 2014. Encyclopedia of language development. Lexical Innovations.

Eve V. Clark. 1980. Lexical innovations: How children learn to create new words. papers and reports on child language development, number 18. *null*.

Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*.

Herbert H Clark. 1998. 4 communal lexicons.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Fried, Justin Chiu, and Dan Klein. 2021. Reference-centric models for grounded collaborative dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2130–2147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment.

Baber Khalid, Malihe Alikhani, Michael Fellner, Brian McMahan, and Matthew Stone. 2020a. Discourse coherence, reference grounding and goal oriented dialogue. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtually at Brandeis, Waltham, New Jersey. SEMDIAL.

Baber Khalid, Malihe Alikhani, and Matthew Stone. 2020b. Combining cognitive modeling and reinforcement learning for clarification in dialogue. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4417–4428, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hyounghun Kim, Hao Tan, and Mohit Bansal. 2020. Modality-balanced models for visual dialogue.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Ernest Lepore, Ernest Lepore, and Matthew Stone. 2010. Against metaphorical meaning. *Topoi-an International Review of Philosophy*.

David Lewis. 2008. *Convention: A philosophical study*. John Wiley & Sons.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco:

Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea. Association for Computational Linguistics.

Changsong Liu, Rui Fang, Lanbo She, and Joyce Chai. 2013. Modeling collaborative referring for situated referential grounding. In *Proceedings of the SIGDIAL 2013 Conference*, pages 78–86, Metz, France. Association for Computational Linguistics.

Matthew Marge, Carol Y. Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gilmer L. Blankenship, Joyce Chai, Hal Daumé III, Debadeepta Dey, Mary P. Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David R. Traum, and Zhou Yu. 2022. Spoken language interaction with robots: Recommendations for future research. *Comput. Speech Lang.*, 71:101255.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*.

Will S. Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Teach: Task-driven embodied agents that chat.

David Schlangen. 2019. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings.

Takashi Sugahara. 2011. Onomatopoeia in spoken and written english: Corpus- and usage-based analysis.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Alberto Testoni, Claudio Greco, and Raffaella Bernardi. 2022. Artificial intelligence models do not ground negation, humans do. guesswhat?! dialogues as a case study. *Frontiers in Big Data*, 4.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363.

Olivia Winn and Smaranda Muresan. 2018. 'lighter' can still be dark: Modeling comparative color descriptions. *ACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.