UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Refer, Reuse, Reduce

Generating Subsequent References in Visual and Conversational Contexts

Takmaz, E.; Giulianelli, M.; Pezzelle, S.; Sinclair, A.; Fernández, R.

DOI
10.18653/v1/2020.emnlp-main.353

Publication date
2020

Document Version
Final published version

Published in
2020 Conference on Empirical Methods in Natural Language Processing

License
CC BY

Link to publication

Citation for published version (APA):
Takmaz, E., Giulianelli, M., Pezzelle, S., Sinclair, A., & Fernández, R. (2020). *Refer, Reuse, Reduce*: Generating Subsequent References in Visual and Conversational Contexts. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *2020 Conference on Empirical Methods in Natural Language Processing: EMNLP 2020 : proceedings of the conference : November 16-20, 2020* (pp. 4350-4368). The Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.353

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Oct 2026

Refer, Reuse, Reduce
Generating Subsequent References in Visual and Conversational Contexts

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{e.takmaz|m.giulianelli|s.pezzelle|
a.j.sinclair|raquel.fernandez}@uva.nl

Abstract

Dialogue participants often refer to entities or situations repeatedly within a conversation, which contributes to its cohesiveness. Subsequent references exploit the common ground accumulated by the interlocutors and hence have several interesting properties, namely, they tend to be shorter and reuse expressions that were effective in previous mentions. In this paper, we tackle the generation of first and subsequent references in visually grounded dialogue. We propose a generation model that produces referring utterances grounded in both the visual and the conversational context. To assess the referring effectiveness of its output, we also implement a reference resolution system. Our experiments and analyses show that the model produces better, more effective referring utterances than a model not grounded in the dialogue context, and generates subsequent references that exhibit linguistic patterns akin to humans.

1 Introduction

When speakers engage in conversation, they often refer to the same objects or situations more than once. Subsequent references (McDonald, 1978) are dependent on the shared knowledge that speakers accumulate during dialogue. For example, dialogue participants may first mention *"a white fuzzy dog with a wine glass up to his face"* and later refer to it as *"the wine glass dog"*, as shown in Figure 1, dialogue 1. Speakers establish 'conceptual pacts', i.e., particular ways of conceptualising referents that condition what is perceived as coherent in a given dialogue (Garrod and Anderson, 1987; Brennan and Clark, 1996). While *"the wine glass dog"* may be odd as a standalone description, it is an appropriate referring expression in the above conversational context. Yet, uttering it in a different context (such as dialogue 2 in Figure 1, after the participants had successfully referred to the image as *"the dog on the red chair"*) may disrupt the cohesion of the dialogue and lead to communication problems (Metzing and Brennan, 2003).

Referring utterances extracted from dialogue 1
A: a white fuzzy dog with a wine glass up to his face
⇝ B: I see the wine glass dog
⇝ A: no I don't have the wine glass dog

Referring utterances extracted from dialogue 2
C: white dog sitting on something red
⇝ D: yes I have the dog on the red chair
⇝ C: white dog on the red chair

Figure 1: Two chains of referring utterances from two games with different participants, including the first description of the target image in that dialogue and two subsequent references (⇝). In the game, each participant sees 5 additional images besides the target shown here. The distractor images change at every round of the game, i.e., each co-referring utterance within a dialogue is produced in a slightly different visual context.

In this paper, we tackle the generation of referring utterances—i.e., utterances that contain referring descriptions, as in Figure 1—grounded both in the visual environment and the dialogue context. These utterances have several interesting properties that make their automatic generation challenging. First, they are produced with the communicative goal of helping the addressee identify the intended referent. Second, because humans operate under cognitive and time-bound constraints, dialogue participants will aim to fulfil this communicative goal while optimising the use of their limited cognitive resources. This results in two common fea-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4350–4368,
November 16–20, 2020. ©2020 Association for Computational Linguistics

tures of subsequent mentions: (1) *Reduction:* Utterances tend to become shorter—a well attested phenomenon since the work of Krauss and Weinheimer (1967)—as a result of interlocutors' reliance on their common ground (Stalnaker, 2002): As more shared information is accumulated, it becomes predictable and can be left implicit (Grice, 1975; Clark and Wilkes-Gibbs, 1986; Clark and Brennan, 1991; Clark, 1996). Sentence compression also takes place in discourse, as predicted by the entropy rate principle (Genzel and Charniak, 2002; Keller, 2004). (2) *Lexical entrainment:* Speakers tend to reuse words that were effective in previous mentions (Garrod and Anderson, 1987; Brennan and Clark, 1996) possibly due to priming effects (Pickering and Garrod, 2004). Thus, besides being a challenging problem intriguing from a linguistic and psycholinguistic point of view, computationally modelling the generation of subsequent references can contribute to better user adaptation in dialogue systems and to more natural human-computer interaction.

For our study, we use data from the PhotoBook dataset (Haber et al., 2019), developed to elicit subsequent references to the same images within task-oriented dialogue. To isolate the issue we are interested in, we extract, from each dialogue, the utterances that refer to a given image. This results in a dataset of dialogue-specific chains of co-referring utterances: For example, Figure 1 shows two chains of co-referring utterances from two different dialogues, both referring to the same image. Figure 2 shows another example. We then formulate the problem as the generation of the next utterance in a chain given the current visual context and the common ground established in previous co-referring utterances (whenever these are available). To computationally model this problem, we propose three variants of a generation system based on the encoder-decoder architecture (Sutskever et al., 2014). We evaluate their output with metrics commonly used in the domain of Natural Language Generation and with several linguistic measures. In addition, to assess the communicative effectiveness of the generated references, we implement a reference resolution agent in the role of addressee.

We find that conditioning the generation of referring utterances on previous mentions leads to better, more effective descriptions than those generated by a model that does not exploit the conversational history. Furthermore, our quantitative and qualitative analysis shows that the context-aware model generates subsequent references that exhibit linguistic patterns akin to humans' regarding markers of new vs. given information, reduction, and lexical entrainment, including novel noun-noun compounds.

Our data, code, and models are available at `https://dmg-photobook.github.io`.

2 Related Work

Generation of distinguishing expressions Our work is related to Referring Expression Generation (REG), a task with a long tradition in computational linguistics that consists in generating a description that distinguishes a target from a set of distractors—Krahmer and van Deemter (2012) provide an overview of early approaches. Follow-up approaches focused on more data-driven algorithms exploiting datasets of simple visual scenes annotated with symbolic attributes (e.g., Mitchell et al., 2013a,b, among others). More recently, the release of large-scale datasets with real images (Kazemzadeh et al., 2014) has made it possible to test deep learning multimodal models on REG, sometimes in combination with referring expression comprehension (Mao et al., 2016; Yu et al., 2017). While REG typically focuses on describing objects within a scene, a few approaches at the intersection of REG and image captioning (Bernardi et al., 2016) have aimed to generate discriminative descriptions of full images, i.e., image captions that can distinguish the target image from a pool of related ones (Andreas and Klein, 2016; Vedantam et al., 2017; Cohn-Gordon et al., 2018). Similarly to these approaches, in the present work, we generate utterances that refer to a full image with the aim of distinguishing it from other distractor images. In addition, our setup has several novel aspects: The referring utterances are the result of interactive dialogue between two participants and include subsequent references.

Generation of subsequent references Follow-up work within the REG tradition has extended the early algorithms to deal with subsequent references (Gupta and Stent, 2005; Jordan and Walker, 2005; Stoia et al., 2006; Viethen et al., 2011). These approaches focus on content selection (i.e., on generating a list of attribute types such as `color` or `kind` using an annotated corpus) or on choosing the type of reference (definite or indefinite noun phrase, pronoun, etc.) and do not directly exploit visual representations. In contrast,

we generate the surface realisation of first and subsequent referring utterances end-to-end, grounding them in continuous visual features of real images.

Our work is related to a recent line of research on reference *resolution* in visually-grounded dialogue, where previous mentions have been shown to be useful (Shore and Skantze, 2018; Haber et al., 2019; Roy et al., 2019). Here we focus on *generation*. To our knowledge, this is the first attempt at generating visually grounded referring utterances taking into account earlier mentions in the dialogue. Some work on generation has exploited dialogue history in order to make lexical choice decisions that align with what was said before (Brockmann et al., 2005; Buschmeier et al., 2009; Stoyanchev and Stent, 2009; Lopes et al., 2015; Hu et al., 2016; Dušek and Jurčíček, 2016). Indeed, incorporating entrainment in dialogue systems leads to an increase in the perceived naturalness of the system responses and to higher task success (Lopes et al., 2015; Hu et al., 2016). As we shall see, our generation model exhibits some lexical entrainment.

Dialogue history in visual dialogue Recent work in the domain of visually grounded dialogue has exploited dialogue history in encoder-decoder models trained on large datasets of question-answering dialogues (Das et al., 2017; De Vries et al., 2017; Chattopadhyay et al., 2017). Recently, Agarwal et al. (2020) showed that only 10% of the questions in the VisDial dataset (Das et al., 2017) genuinely require dialogue history in order to be answered correctly, which is in line with other shortcomings highlighted by Massiceti et al. (2018). More generally, visually grounded dialogue datasets made up of sequences of questions and answers lack many of the collaborative aspects that are found in natural dialogue. For our study, we focus on the PhotoBook dataset by Haber et al. (2019), where dialogues are less restricted and where the common ground accumulated over the dialogue history plays an important role.

3 Data

3.1 PhotoBook Dataset

The PhotoBook dataset (Haber et al., 2019) is a collection of task-oriented visually grounded English dialogues between two participants. The task is set up as a game comprised of 5 rounds. In each round, the two players are assigned private 'photo books' of 6 images, with some of those images being present in both photo books. The goal is to find out which images are common to both players by interacting freely using a chat interface. In each round, the set of 6 images available to each player changes, but a subset of images reappears, thus triggering subsequent references to previously described images. This feature of the PhotoBook dataset makes it a valuable resource to model the development of conversational common ground between interlocutors. The dataset consists of 2,500 games, 165K utterances in total, and 360 unique images from MS COCO (Lin et al., 2014).

3.2 Dataset of Referring Utterance Chains

As mentioned above, in PhotoBook participants can freely interact via chat. The dialogues thus include different types of dialogue act besides referring utterances. While utterances performing other functions are key to the dialogue and may provide useful information, in the present work we abstract away from this aspect and concentrate on referring utterances.[1] To create the data for our generation task, we extract utterances that contain an image description and their corresponding image target from the dialogues as follows. Within a game round, we consider all the utterances up to the point where a given image i has been identified by the participants[2] as candidate referring utterances for i – see Figure 2. We then compare each candidate against a reference set of descriptions made up of the MS COCO (Lin et al., 2014) captions for i and the attributes and relationship tokens of i in the Visual Genome (Krishna et al., 2017). We score each candidate utterance with the sum of its BERTScore[3] (Zhang et al., 2020) for captions and its METEOR score (Banerjee and Lavie, 2005) for attributes and relationships. The top-scoring utterance in the game round is selected as a referring utterance for i and used as an additional caption for extracting subsequent references in the following game rounds. As a result of this procedure, for a given dialogue and an image i, we obtain a reference chain made up of the referring utterances—maximum one per round—that refer to i in the dialogue. Since images do not always reappear in each round, chains can have different

[1] Haber et al. (2019) extracted co-reference chains made up of multi-utterance dialogue excerpts. Our chains include single utterances, which is more suitable for generation.

[2] Image identification actions are part of the metadata.

[3] BERTScore uses contextualised embeddings (Devlin et al., 2019) to assess similarity between a target sentence and one or more reference sentences.

DIALOGUE FRAGMENT AND IMAGES VISIBLE TO PARTICIPANT A IN THE FIRST ROUND OF A GAME

A: Hi
B: Hello.
B: do you have a white cake on multi colored striped cloth?
A: **I see a guy taking a picture. What about you?**
B: is it of a cake with construction trucks on it?
A: Yeah. I don't see the cake you mentioned.
A: `<common img_4>`

RESULTING REFERRING UTTERANCE CHAIN WITH SUBSEQUENT REFERENCES EXTRACTED FROM THE FOLLOWING GAME ROUNDS

1. I see a guy taking a picture. What about you?
2. guy with camera
3. I have the guy with camera
4. The last one is the camera guy.

Figure 2: Example from our new dataset of referring utterance chains. Given a target image selected by a participant (here `<common img_4>`), the utterances in the dialogue prior to that selection action are scored by their likelihood of referring to the target. In this example, the utterance in bold is selected as the first description. To construct the reference chain, subsequent references are extracted in a similar manner from the dialogue in the following game rounds. The set of distractor images available to a participant changes across rounds.

length. Two examples of chains of length 3 are shown in Figure 1 and a chain of length 4 in Figure 2. Given that each utterance in a chain belongs to a different game round, each utterance was produced in a slightly different visual context with different distractor images. Figure 2 shows the visual context available to participant A in the first round of a game, when the participant produced the first description in the dialogue for target image number 4. The other three descriptions in the chain were produced while seeing different distractors.

We evaluate the referring utterance extraction procedure and the resulting chains using 20 dialogues hand-annotated by Haber et al. (2019) with labels linking utterances to the target image they describe. Using our best setup, we obtain a precision of 0.86 and a recall of 0.61. The extracted chains are very similar to the human-annotated ones in terms of chain and utterance length.

Our new dataset is made up of 41,340 referring utterances and 16,525 chains (i.e., there are 16,525 first descriptions and 24,815 subsequent references). The median number of utterances in a chain is 3. We use the splits defined by Haber et al. (2019) to divide the dataset into Train, Validation, and Test, and all hand-annotated dialogues are excluded from these splits. Table 1 reports relevant descriptive statistics of the dataset. More details about the extraction procedure and the dataset are available in Appendix A. Appendix B describes how the dataset is further processed to be used in our models.

Split	Games	*First*		*Later*	
		N	Length	N	Length
Train	1725	11540	10.52 (4.80)	17393	7.52 (4.15)
Val	373	2503	10.49 (4.81)	3749	7.70 (4.22)
Test	368	2482	10.52 (4.85)	3673	7.59 (4.17)

Table 1: Number of games and referring utterances in the splits of our dataset with their average length in tokens (standard deviation in brackets) broken down by first mentions vs. subsequent ('Later') references.

4 Models

With the new dataset of referring utterance chains in place, we operationalise the problem of generating a referring utterance taking into account the visual and conversational context as follows. The model aims to generate a referring utterance given (a) the *visual context* in the current game round made up of 6 images from the perspective of the player who produced the utterance, (b) the *target* among those images, and (c) the *previous co-referring utterances* in the chain (if any). Besides being contextually appropriate, the generated utterance has to be informative and discriminative enough to allow an addressee to identify the target image. We thus also develop a reference resolution model that plays the role of addressee. The two models are trained independently.

4.1 Generation Models

We propose three versions of the generation model, which all follow the encoder-decoder architecture (Sutskever et al., 2014). These versions differ from each other with respect to whether and how

they exploit earlier referring utterances for the target image: (1) a baseline model that does not use the dialogue context at all (hence, **Ref**); (2) a model that conditions the generation on the previous referring utterance, if available, and operates attention over it (hence, **ReRef**); (3) a model that builds on (2) by adding a 'copy' mechanism (See et al., 2017) (hence, **Copy**). We describe them below and provide further details in Appendix C.

Ref This model is provided only with the information about the visual context in the current game round–and not with the linguistic context in previous rounds. We encode each image in the context by means of visual features extracted from the penultimate layer of ResNet-152 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009). First, the visual features of the 6 candidate images are concatenated. This concatenated vector goes through dropout, a linear layer and ReLU (Nair and Hinton, 2010). The same process is applied for the single target image. We then concatenate the final visual context vector with the target image vector, apply a linear transformation, and use the resulting hidden representation h_d to initialise an LSTM decoder, which generates the referring utterance one word at a time. At each timestep, the input to the decoder is a multimodal vector, i.e., the concatenation of h_d and the word embedding of token t_t. The weights of the embeddings are initialised uniformly in the range $(-0.1, 0.1)$ and learned from scratch for the task at hand.

ReRef With this model, we aim to simulate a speaker who is able to *re-refer* to a target image in accordance with what has been established in the conversational common ground (Clark, 1996; Brennan and Clark, 1996). The model enriches **Ref** by incorporating linguistic information into the encoder (in addition to visual information) and an attention mechanism applied over the hidden states of the encoder during decoding. The model thus generates a new utterance conditioned on both the visual and the linguistic context.

The encoder is a one-layer bidirectional LSTM initialised with the same visual input fed to **Ref**. In addition, it receives as input the previous referring utterance used in the dialogue to refer to the target image,[4] or else is fed the special <nohs> token, indicating that there is no conversational history for the target image yet. We utilise the attention mechanism proposed by Bahdanau et al. (2018) and used by See et al. (2017). During decoding, attention contributes to determining which aspects of the multimodal context are most critical in generating the next referring utterance. We expect this attention mechanism to be able to identify the words in a previous utterance that should be present in a subsequent reference, resulting in lexical entrainment.

Copy This model builds on **ReRef** and incorporates a means of simulating lexical entrainment more explicitly, by regulating when a word used in the previous mention should be used again in the current referring utterance (i.e., should be produced by the decoder). Given the shortening property of subsequent references mentioned in the Introduction, our task bears some similarity to text summarisation. We thus draw inspiration from the summarisation model proposed by See et al. (2017). In particular, we equip the model with their 'copy' mechanism, which combines the probability of copying a word present in the encoded input with the probability of generating that word from the vocabulary. We expect this mechanism to contribute to generating rare words present in preceding referring utterances that are part of a 'conceptual pact' (Brennan and Clark, 1996) between the dialogue participants, but may have low generation probability overall.

4.2 Reference Resolution Model

Given an utterance referring to a target image and a 6-image visual context, our reference resolution model predicts the target image among the candidates. This model is similar to the resolution model proposed by Haber et al. (2019) for the PhotoBook dataset, but includes several extensions: (1) We use BERT embeddings from the uncased base BERT model (Devlin et al., 2019; Wolf et al., 2019) to represent the linguistic input rather than LSTMs;[5] (2) The input utterance is encoded taking into account the visual context: We compute a multimodal representation of the utterance by concatenating each BERT token representation with the visual context representation, obtained in the same way as for the generation models;[6] (3) We apply at-

[4] The latest description seems to contain the most relevant information. Including all referring utterances in the chain up to that point in the dialogue did not lead to improvements.

[5] In the generation models, we did not use BERT due to the difficulties of using contextualised embeddings in the decoder, and the desirability of using the same word embeddings in both the encoder and the decoder.

[6] We also tried using multimodal representations obtained via LXMERT (Tan and Bansal, 2019). No improvements were

tention over the multimodal representations of the utterance in the encoder instead of using the output from a language-only LSTM encoder. The utterance's final representation is given by the weighted average of these multimodal representations with respect to the attention weights.

Each candidate image is represented by its ResNet-152 features (He et al., 2016) or, if it has been previously referred to in the dialogue, by the sum of the visual features and the representation of the previous utterance (obtained via averaging its BERT embeddings).[7] To pick a referent, we take the dot product between the representation of the input utterance and each of the candidate image representations. The image with the highest dot-product value is the one chosen by the model.

4.3 Model Configurations

For each model, we performed hyperparameter search for batch size, learning rate, and dropout; also, the search included different dimensions for the embedding, attention, and hidden layers. All models were trained for up to 100 epochs (with a patience of 50 epochs in the case of no improvement to the validation performance) using the Adam optimiser (Kingma and Ba, 2015) to minimise the Cross Entropy Loss with sum reduction. BERTScore F1 (Zhang et al., 2020) in the validation set was used to select the best model for the generation task, while we used accuracy for the resolution task. In the next section, we report average scores and standard deviations over 5 runs with different random seeds. Further details on hyperparameter selection, model configurations, and reproducibility can be found in Appendix E.

5 Results

5.1 Evaluation Measures

We evaluate the performance of the reference resolution model by means of both accuracy and Mean Reciprocal Rank (MRR). As for the generation models, we compute several metrics that are commonly used in the domain of Natural Language Generation. In particular, we consider three measures based on n-gram matching: BLEU-2 (Papineni et al., 2002),[8] ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015). We also compute BERTScore F1 (Zhang et al., 2020) (used for model selection), which in our setup compares the contextual embeddings of the generated sentence to those of the set of referring utterances in the given *chain*. Further details of the metrics are in Appendix D.

All these measures capture the degree of similarity between generated referring utterances and their human counterparts. In addition, to assess the extent to which the generated utterances fulfil their communicative goal, we pass them to our reference resolution model and obtain accuracy and MRR. While this is not a substitute for human evaluation, we take it to be an informative proxy. In Section 6, we analyse the generated utterances with respect to linguistic properties related to phenomena that are not captured by any of these metrics.

5.2 Reference Resolution Results

Our reference resolution model achieves an accuracy of 85.32% and MRR of 91.20% on average over 5 runs. This is a substantial result. A model that predicts targets at random would yield an accuracy of roughly 16.67% (as the task is to pick one image out of 6 candidates), while a baseline that simply takes one-hot representations of the image IDs in the context achieves 22.37% accuracy.[9]

Subset	ACC	MRR	Instances
First	80.27 (0.46)	87.78 (0.28)	2482
Later	88.74 (0.18)	93.51 (0.09)	3673
Overall	85.32 (0.19)	91.20 (0.10)	6155

Table 2: Test set scores of the reference resolution model: averages of 5 runs with the best configuration, with the standard deviations in parentheses.

In Table 2, the results are presented by breaking down the test set into two subsets: the *first* referring utterances in a chain, and *later* referring utterances, i.e., subsequent references where the target image among the candidates has linguistic history associated with it. The model performs better on subsequent references. Exploiting dialogue history plays a role in this boost: an ablated version of the model that does not have access to the linguistic history of subsequent references yields an accuracy of 84.82% for the *Later* subset, which is significantly lower than the 88.74% obtained with

observed.

[7] Thus, some of the candidate images have multimodal representations (if they were already mentioned in the dialogue), while others do not.

[8] BLEU-2, which is based on bigrams, appears to be more informative than BLEU with longer n-grams in dialogue response generation (Liu et al., 2016)

[9] In this simple baseline, one-hot vectors are projected to scalar values, and a softmax layer assigns probabilities over them. The fact that this is slightly higher than random accuracy seems due the different frequencies of images being the target.

our model ($p < 0.01$ independent samples t-test). This confirms the importance of accessing information about previous mentions in visually grounded reference resolution (Haber et al., 2019).

We use the best model run to assess the communicative effectiveness of our generation models.

5.3 Generation Model Results

As we did for the reference resolution model, we break down the test set into first referring utterances in a chain and subsequent references, for which generation is conditioned on a previous utterance. The outcomes of this breakdown are provided in Table 3, where we report the test set performances of our three generation models. Overall results on the validation set are available in Appendix F.

ReRef obtains the highest scores across all measures, followed by **Copy**, while **Ref** achieves substantially lower results. Regarding the comparison between first and subsequent references, the context-aware models **ReRef** and **Copy** attain significantly higher results when generating later mentions vs. first descriptions ($p < 0.001$, independent samples t-test). As expected, no significant differences are observed in this respect for **Ref**.[10]

As for the communicative effectiveness of the generated utterances as measured by our resolution model, both accuracy and MRR are particularly high (over 90%) for **ReRef**. Across all model types, generated subsequent references are easier to resolve by the model, in line with the pattern observed in Table 2 for the human data.

All in all, the addition of the copy mechanism does not provide improvements over **ReRef**'s performance that can be detected with the current evaluation measures. We do find, however, that the **Copy** model uses a substantially larger vocabulary than **ReRef**: 1,791 word types vs. 760 (the human vocabulary size on the test set is 2,332, while **Ref** only uses 366 word types). An inspection of the vocabularies shows that **Copy** does generate a good deal of low-frequency words, in line with what is expected from the dedicated copy mechanism (less desirably, this also includes words with spelling errors). Further analysis also shows that **Copy** generates utterances that include more repetitions: 18% of the utterances generated by **Copy** in the test set contain two identical content words e.g. *"do you have the runway runway woman?"*, while only 7% of those generated by **ReRef** do.[11] Adding a means to control for repetitions, such as the 'coverage' mechanism by See et al. (2017), could be worth exploring in the future.

We compare our best performing model **ReRef** to a baseline consisting in reusing the first generated utterance verbatim in later mentions. In this case, the model does not learn how to reuse previous referring utterances taking into account the changing visual context, but simply keeps repeating the first description it has generated. We expect this baseline to be relatively strong given that experiments in the lab have shown that dialogue participants may stick to an agreed description even when some properties are not strictly needed to distinguish the referent in a new visual context (Brennan and Clark, 1996; Brown-Schmidt et al., 2015). The results (reported in Table 3 *baseline*) show that the model significantly outperforms this baseline when generating later mentions.

Overall, our results confirm that referring utterances do evolve during a dialogue and indicate that the models that exploit the conversational context are able to learn some of the subtle modifications involved in the re-referring process. In the next section, we look into the linguistic patterns that characterise this process.

6 Linguistic Analysis

We analyse the linguistic properties of the utterances generated by the best performing run of each of our models and compare them with patterns observed in the human data. Extensive descriptive statistics are available in Appendix G.

6.1 Main Trends

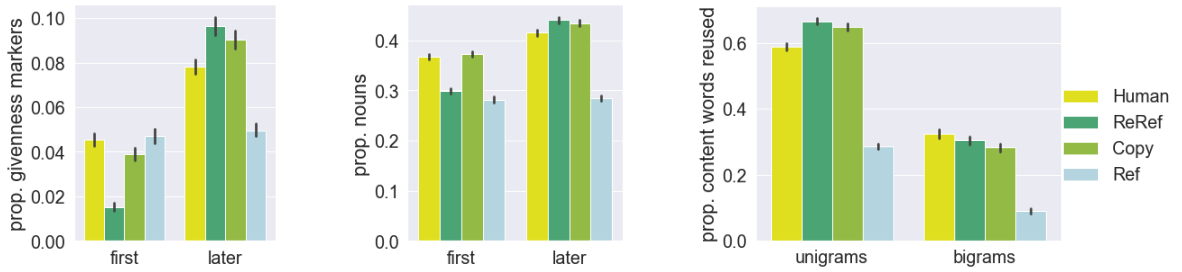
Givenness markers We first look into the use of markers of new vs. given information, in particular indefinite and definite articles as well as particles such as *again* or *before* (as in *"I have the X one again"* or *"the X from before"*), which are anaphoric and presuppose that an image has been discussed previously in the dialogue. Figure 3a shows the proportion of givenness markers (*the, one, same, again, also, before*) in first vs. subsequent references. Not surprisingly, this proportion increases in the human subsequent references. **ReRef** and **Copy** both display an amplified version

[10] While first descriptions do not require linguistic context, **ReRef** and **Copy** perform better on first description generation than **Ref**. This is likely due to their higher complexity.

[11] The **Ref** model is even more repetitive: 21% of the generated utterances contain repeated content words.

Model	Subset	BLEU-2	ROUGE	CIDEr	BERT-F1	ACC	MRR
Ref	**First**	20.80 (1.02)	29.74 (1.59)	41.26 (3.14)	54.48 (1.38)	57.12 (4.85)	72.47 (3.19)
	Later	23.06 (1.20)	31.88 (1.66)	40.79 (2.83)	55.54 (1.40)	60.94 (2.67)	75.34 (1.59)
ReRef	**First**	33.09 (0.79)	42.32 (0.42)	94.63 (2.12)	62.55 (0.12)	90.36 (1.73)	94.49 (1.14)
	Later	52.15 (1.19)	56.74 (0.63)	143.59 (5.84)	71.25 (0.39)	92.21 (0.73)	95.62 (0.45)
	baseline	36.66 (0.92)	45.37 (0.57)	96.41 (2.69)	64.13 (0.24)	90.14 (2.28)	94.38 (1.41)
Copy	**First**	25.25 (0.40)	33.31 (0.50)	60.51 (1.21)	57.61 (0.36)	81.36 (0.53)	88.70 (0.49)
	Later	43.08 (0.36)	48.79 (0.41)	128.45 (1.98)	66.07 (0.17)	83.96 (0.53)	90.60 (0.32)

Table 3: Test set scores of the generation models (averaged over 5 runs) for first vs. subsequent references, including word-overlap metrics, BERTScore F1, and accuracy/MRR obtained by our resolution model on the generated utterances. ReRef *baseline* uses the first generated description verbatim in all later mentions. All differences across model types are statistically significant ($p < 0.001$, independent samples t-test).



(a) Givenness markers (b) Proportion of nouns (c) Lexical entrainment in later references

Figure 3: Linguistic patterns in human referring utterances and in referring utterances generated by our three models. Givenness markers and proportion of nouns per utterance are displayed for first and later references.

of this trend, while **Ref**, which cannot capture any given information, shows no difference.

Reduction Regarding referring utterance length, we observe a significant shortening in subsequent mentions in human dialogues (11.3 vs. 8.3 tokens on average in first and subsequent mentions, respectively). This shortening is also observed in the utterances generated by **ReRef** (11.3 vs. 7.2) and **Copy** (10.8 vs. 7.8). **Ref** tends to generate longer utterances across the board (13.7 vs. 13.6).

Shortening may be linked to compression, i.e., to an increase in information density (Shannon, 1948). To analyse this aspect, we consider the proportion of content words in the utterances, since such proportion can capture mechanisms such as syntactic reduction (e.g., the removal of the complementiser *that*), which has been shown to be a good predictor of information density increase (Levy and Jaeger, 2006). Haber et al. (2019) reported a rise in the proportion of content words for all utterance types in later rounds of the PhotoBook games. We also observe such an increase in our referring utterance chains, and a similar trend is exhibited as well by the output of the **ReRef** and **Copy** models: In particular, generated subsequent references contain a significantly higher proportion of nouns and adjectives compared to first descriptions. Figure 3b shows this pattern for nouns, which are the most prominent type of content word in our data.

Entrainment In order to analyse the presence of lexical entrainment, we compute the proportion of expressions in subsequent references that are reused from the previous mention. We compare reuse at the level of unigrams and bigrams. Figure 3c shows this information focusing on content words. Around 60% of content tokens are reused by humans. The proportion is even higher in the utterances generated by our context-aware models. Digging deeper into the types of content tokens being reused, we find that nouns are reused significantly more than other parts of speech by humans. This is also the case in the subsequent references generated by the **ReRef** and **Copy** models.

Humans also reuse a substantial proportion of content word bigrams—as do, to a smaller degree, the context-aware models. For example, given the gold description *"pink bowls rice and broccoli salad next to it"*, **ReRef** generates the subsequent reference *"pink bowls again"*. Noun-noun compounds are a particularly interesting case of such

bigrams, which we qualitatively analyse below.

6.2 A Case Study: Noun-Noun Compounds

A partial manual inspection of the human utterances in our chains reveals that, as they proceed in the dialogue, participants tend to produce referring expressions consisting of a noun-noun compound.[12] For example, in Figure 2 we observe the compound *"camera guy"* being uttered after the previous mention *"guy with camera"*. (reused nouns are underlined). Another example is *"wine glass dog"* in Figure 1. This is in line with Downing (1977), who argues that novel (i.e., not yet lexicalised) noun-noun compounds can be built by speakers on the fly based on a temporary, implicit relationship tying the two nouns, e.g., 'the guy *taking a picture with* a camera'. Such noun-noun compounds are thus prototypical examples of reuse and reduction: On the one hand, the novel interpretation (which needs to be pragmatically informative, diagnostic, and plausible; Costello and Keane, 2000) can only arise from the established common ground between speakers; on the other hand, compounds are naturally shorter than the 'source' expression since they leave implicit the relation between the nouns.

We check whether our best performing generation models produce compounds as humans do, i.e., by reusing nouns that were previously mentioned while compressing the sentence. We perform the analysis with a qualitative focus, by manually inspecting a subset of the generated utterances.[13] In Figure 4, we show two noun-noun compounds generated by **ReRef** (similar cases were observed for **Copy**). The example on the left is a noun-noun compound, *"basket lady"*, that is consistent with the dialogue context: both nouns are indeed reused from the previous mention. In contrast, the compound on the right does not build on the conversational history; the noun *"tattoo"* is not in the previous mention and never uttered within the reference chain (not reported), and thus may be perceived as breaking a conceptual pact (Metzing and Brennan, 2003). The compound is grounded in the image, but not in the conversational context.

[12] This is consistent with the fact that the proportion of noun-noun bigrams is significantly higher in subsequent references (0.05 vs. 0.08 on average in first and subsequent references, respectively; $p < 0.001$ independent sample t-test).

[13] The subset is obtained by applying simple heuristics to the set of generated utterances, such as length and PoS tags.

P: lady with basket?
⇝ **ReRef:** basket lady?

P: do you have headband guy?
⇝ **ReRef:** tattoo guy?

Figure 4: Two examples from the test set where **ReRef** generates a noun-noun compound based on the previous human mention (***P***). Left: a genuine *reuse* case; right: a *non-reuse* case. Reused words are underlined.

7 Conclusion

We have addressed the generation of descriptions that are (1) discriminative with respect to the visual context and (2) grounded in the linguistic common ground established in previous mentions. To our knowledge, this is the first attempt at tackling this problem at the level of surface realisation within a multimodal dialogue context.

We proposed an encoder-decoder model that is able to generate both first mentions and subsequent references by encoding the dialogue context in a multimodal fashion and dynamically attending over it. We showed that our best performing model is able to produce better, more effective referring utterances than a variant that is solely grounded in the visual context. Our analysis revealed that the generated utterances exhibit linguistic properties that are similar to those observed in the human utterances regarding reuse of words and reduction. Generating subsequent references with such properties has the potential to enhance user adaptation and successful communication in dialogue systems.

Yet, in our approach we abstracted away from important interactive aspects such as the collaborative nature of referring in dialogue (Clark and Wilkes-Gibbs, 1986), which was considered by Shore and Skantze (2018) for the task of reference resolution. In the present work, we simplified the interactive aspects of reference by extracting referring utterances from the PhotoBook dialogues and framing the problem as that of generating the next referring utterance given the previous mention. We believe that the resulting dataset of referring utterance chains can be a useful resource to analyse and model other dialogue phenomena, such as saliency or partner specificity, both on language alone or on the interaction of language and vision.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

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*. To appear.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Pushmeet Kohli, and Edward Grefenstette. 2018. Jointly learning "what" and "how" from instructions and goal-states. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Carsten Brockmann, Amy Isard, Jon Oberlander, and Michael White. 2005. Modelling alignment for affective dialogue. In *Workshop on adapting the interaction style to affective factors at the 10th international conference on user modeling (UM-05)*.

Sarah Brown-Schmidt, Si On Yoon, and Rachel Anna Ryskin. 2015. People as contexts in conversation. In *Psychology of Learning and Motivation*, volume 62, chapter 3, pages 59–99. Elsevier.

Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2009. An alignment-capable microplanner for Natural Language Generation. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 82–89, Athens, Greece. Association for Computational Linguistics.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-AI games. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1 – 39.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Fintan J Costello and Mark T Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

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.

Pamela Downing. 1977. On the creation and use of english compound nouns. *Language*, 53(4):810–842.

Ondřej Dušek and Filip Jurčíček. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 185–190, Los Angeles. Association for Computational Linguistics.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181 – 218.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

H. Paul Grice. 1975. Logic and conversation. In D. Davidson and G. Harman, editors, *The Logic of Grammar*, pages 64–75. Dickenson, Encino, California.

Surabhi Gupta and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, pages 1–6.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Zhichao Hu, Gabrielle Halberg, Carolynn R Jimenez, and Marilyn A Walker. 2016. Entrainment in pedestrian direction giving: How many kinds of entrainment? In *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 151–164. Springer.

Pamela W Jordan and Marilyn A Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona, Spain. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Robert M. Krauss and Sidney Weinheimer. 1967. Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning & Verbal Behavior*, 6(3):359–363.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting Language and Vision using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73.

Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 849–856. MIT Press.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, pages 740–755. Springer.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Daniela Massiceti, Puneet K. Dokania, N Siddharth, and Philip H. S. Torr. 2018. Visual dialogue without vision or dialogue. In *NeurIPS Workshop On Critiquing And Correcting Trends In Machine Learning*.

David D. McDonald. 1978. Subsequent reference: Syntactic and rhetorical constraints. In *Theoretical Issues in Natural Language Processing-2*.

Charles Metzing and Susan E Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213.

Margaret Mitchell, Ehud Reiter, and Kees Van Deemter. 2013a. Typicality and object reference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. 2013b. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, Madison, WI, USA. Omnipress.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Subhro Roy, Michael Noseworthy, Rohan Paul, Daehyung Park, and Nicholas Roy. 2019. Leveraging past references for robust language grounding. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 430–440, Hong Kong, China. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication, Bell System Technical Journal 27: 379-423 and 623–659. *Mathematical Reviews (MathSciNet): MR10, 133e*.

Todd Shore and Gabriel Skantze. 2018. Using lexical alignment and referring ability to address data sparsity in situated dialog reference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2288–2297, Brussels, Belgium. Association for Computational Linguistics.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 81–88, Sydney, Australia. Association for Computational Linguistics.

Svetlana Stoyanchev and Amanda Stent. 2009. Lexical and syntactic adaptation and their impact in deployed spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 189–192, Boulder, Colorado. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE*

conference on Computer Vision and Pattern Recognition (CVPR), pages 4566–4575.

Jette Viethen, Robert Dale, and Markus Guhe. 2011. Generating subsequent reference in shared visual scenes: Computation vs re-use. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1158–1167, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3521–3529.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Appendices

A Reference chain extraction

For our generation task, we extract reference chains of single referring utterances from the PhotoBook dataset (Haber et al., 2019). Given a dialogue and a target image, a reference chain is comprised of utterances—maximum one per round—that refer to the target image in that dialogue. Due to the size of the PhotoBook dataset (see Section 3.1), we perform this procedure automatically, with a three-step heuristic method described in the following sections. The chain extraction code is available at `https://dmg-photobook.github.io`.

Extracting dialogue segments The goal of segment extraction is to identify all utterances that may include a description of a given target image. To identify relevant segments, we leverage the participants' recorded actions, i.e. selecting an image as common or different (more details on the available metadata in Haber et al., 2019). When an image is selected by a participant as *common* in a dialogue round, we extract all utterances up to that point in the round as candidate referring expressions. We collect referring expressions for a given image in a dialogue starting from the round when *both* speakers observe it. The speakers are then more likely to have established a conceptual pact (see Section 1).

Scoring referring utterances In this second step, we assign a score to each utterance in the extracted segments indicating how likely it is for that utterance to be a description of a given image. To produce these scores, we use as reference the MS COCO image captioning dataset (Lin et al., 2014) and the Visual Genome dataset of scene graphs (Krishna et al., 2017). All 360 pictures in PhotoBook are taken from MS COCO, so we have access to at least 5 captions for each target image. Instead, the Visual Genome dataset provides detailed scene graphs for 37% of the PhotoBook images.

To measure the similarity of a candidate utterance to a reference MS COCO caption, we use the BERTScore (Zhang et al., 2020). We experiment with BERTScore Precision, Recall, F1, and select BERTScore F1. As, in our dialogue setting, utterances often contain lexical material that is not part of a referring expression, we filter out stop-words from both the captions and the utterances. We use spaCy's stop-word list for English from which we remove numerals and prepositions that encode spatial information.[14]. Furthermore, to capture dyad-specific variation in referring language, we add the utterance with the highest BERTScore in a round to the reference set, and use it as an additional caption for the following rounds.

To take into account visual attributes and relationships, for each image we collect attribute tokens $T_A(i)$ (e.g. *leafy*, *tree* from *leafy(tree)*) and relationship tokens $T_R(i)$ (e.g. *man*, *playing*, *frisbee* from *playing(man, frisbee)*) from the Visual Genome dataset of scene graphs. We only consider the intersection $T_{VG}(i) = T_A(i) \cap T_R(i)$ between the sets of attribute and relationship tokens to retain only the most relevant tokens. The set difference $T_{VG}(i*) \setminus \bigcup_{i=1,i\neq i*}^{12}$ between the Visual Genome tokens of the target image and the tokens of the 11 distractors is then used as a reference set. To score an utterance, we compute its METEOR score (Banerjee and Lavie, 2005) with respect to this reference set. For all images annotated in the Visual Genome dataset, the final utterance score is the sum of BERTScore and METEOR.[15]

[14]The English stop-word list is available at `https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py` and our edits at `https://dmg-photobook.github.io`.

[15]We implement BERTScore and use NLTK's code for METEOR (`https://www.nltk.org/api/nltk.translate.html`). We set METEOR's alignment penalty to 0 as our references are unordered collections of tokens.

	Chains	Utterances	Unique utterances	Target images	Image domains	Chain length	Utterance length
Train	11540	28933	27288	360	30	2.51(0.85)	8.71(4.66)
Validation	2503	6252	6009	360	30	2.50 (0.85)	8.82 (4.67)
Test	2482	6155	5876	360	30	2.48 (0.86)	8.77(4.68)
Extracted-20	327	824	807	199	24	2.52 (0.85)	9.50 (4.75)
Gold-20	327	756	740	199	24	2.31 (0.94)	9.47 (4.77)

Table 4: Descriptive statistics of all portions of the extracted dataset of reference utterance chains. Gold-20 is a set of 20 hand-annotated PhotoBook dialogues, with referent labels linking utterances to the target image they describe (see Section 3.2) whereas Extracted-20 are the reference chains extracted from the same 20 dialogues, as if they were not annotated. Duplicate utterances are due to chance: PhotoBook participants have uttered them in different dialogues, potentially to describe the same target image. Image domains refers to the number of MS COCO image categories covered by a dataset portion; the 360 PhotoBook images come from a total of 30 domains.

Selecting referring utterances The last step, utterance selection, produces reference chains consisting of single utterances—maximum one per round. As PhotoBook dialogues are made up of five rounds, reference chains will have a minimum length of 1 and a maximum possible length of 5. First, given an extracted dialogue segment, we discard all utterances produced by speakers who do not have that image in their visual context. Then, for each target image in the corresponding dialogue round, we collect a ranked candidate list of n top-scoring utterances. As an utterance can be selected as a candidate for multiple images in the same round, we discard a candidate *(utterance, image)* pair if its score is lower than that of any other *(utterance', image)* pair in the same round. Finally, we pick the utterance with the highest score among the remaining candidates. For some images, all of the n top-scoring utterances are assigned to other images, and with higher scores. This causes a slight decrease in the number of utterances in the extracted dataset. We set $n = 4$ to minimise the number of discarded utterances. Table 4 reports relevant statistics for the dataset splits of our extracted reference utterance chains.

B Data processing for models

We further process the dataset of automatically extracted utterance chains. Every utterance is uniquely identified by the game ID, round number, message number and the ID of the image that they refer to. From these utterances and their contexts, we build the data we feed into our models.

While providing the 6 candidate images to the reference resolution models, we also keep track of the respective histories of candidates (the last utterance up to that time in the game).

As the distribution of the 6 images and the positions of the target is not uniform for each target-context pair, this may constitute a bias in the reference resolution model. Therefore, to overcome this, we shuffle the images in the context for all splits at the beginning of each epoch. In the generation models, this shuffling is done once at the beginning of training for all splits.

B.1 BERT representations

Since utilising pre-trained BERT models and representations has proven to be beneficial to many NLP tasks (Devlin et al., 2019), we also decided to use BERT to encode the linguistic input in the reference resolution models. For this purpose, we use the BERT-base-uncased model and the tokeniser as provided in the HuggingFace's Transformers library (Wolf et al., 2019). The utterances are first encoded into the correct format for BERT models. Afterwards, they go through the BERT model to produce the hidden states that correspond to the representations of each of the input wordpieces. Finally, all utterances are fed into the reference resolution model in the form of a set of BERT representations.

We also experimented with using BERT-large-uncased model as well as extracting hidden states from multiple layers and aggregating them. Neither option provided further improvements on the results we obtained with the final hidden states from the BERT-base-uncased model. Hence, we opted to use the base model's outputs, where each hidden state is of size 768.

B.2 Embeddings from scratch

For the generation models where we do not use BERT representations, we create a vocabulary of tokens from the training set with the help of TweetTokenizer from the NLTK library[16]. We then map the words that occurred only once in the training

[16] https://www.nltk.org/api/nltk.tokenize.html

split to '<unk>'. This results in a vocabulary of size 2816 (including <pad>, <unk>, <sos>, and <eos>). In addition to these special tokens, we also add <nohs> to point out that there was **no his**tory (no previous utterance) for the target image at that point in the game. This token is utilised in the models that base their generation on the previous utterance. An input of <nohs> means that what the generation model is expected to produce is the very first utterance for that image in the game.

The tokens in all 3 splits are converted to indices using this final vocabulary. For the copy model, we need to keep track of what the actual form of an <unk> token is. For this purpose, we build a full vocabulary from the whole dataset to have access to every word in all splits in their actual surface forms. This vocabulary is of size 5793 (including all 5 special tokens mentioned above).

Since we do not want the generation model to output the <nohs> token, the search space of the decoder does not include this token. The Copy model needs to keep track of unknown tokens in the previous utterance and map the previous utterance using an extended vocabulary so that the decoder would be able to 'copy' from the input itself, rather than only generating words from the reduced vocabulary. Mapped expected next utterance is used in calculating the loss. Actual inputs to the encoder and the decoder still contain unknown words, as we do not maintain special embeddings for the surface forms of each of the unknown tokens.

C Model architectures

Below are more details about our generation models and our reference resolution model.

C.1 Generation models

In these models, we apply teacher forcing during training; therefore, a token embedding at timestep t is the embedding of the expected token from the ground-truth utterance. During validation, the models use the embedding of the word they generated in the previous timestep.

C.1.1 ReRef model

This model obtains the visual input as in the Ref model (consisting of the context and the target). However, instead of initialising the decoder as in the prior model, here, this visual representation initialises the encoder. The encoder receives as input a sentence that was previously used in the same game to refer to the target image (or simply <nohs>, if there was no history for the target image in the game at that point). The embeddings of this input go through dropout.

We concatenate the last hidden states of the forward and backwards directions of the BiLSTM encoder. This concatenated vector is then projected to hidden dimensions and used to initialise the decoder. The input to the decoder during training is an embedding of the ground-truth utterance.

For the attention mechanism, each hidden output of the encoder ${h_{enc}}^t$ (concatenation of forward and backward hidden states for timestep t) goes through a linear layer that projects it from double the size of hidden dimensions to the attention dimensions. In addition, the current hidden state of the decoder ${h_{dec}}^c$ is projected from the hidden dimensions to the attention dimensions.

$$enc^t = W_e {h_{enc}}^t \tag{1}$$

$$dec^c = W_d {h_{dec}}^c \tag{2}$$

$$e_t = v_a(tanh(enc^t + dec^c)) \tag{3}$$

Attention weights are calculated based on the sum of enc^t and dec^c, on which we apply $tanh$ nonlinearity and a linear layer. Padded tokens are masked and softmax is applied over all remaining encoder timesteps i:

$$a_i = softmax(e_i) \tag{4}$$

$$h^* = \sum_i a_i h^i_{enc} \tag{5}$$

To predict the word that the decoder will generate, we concatenate the decoder's current hidden state ${h_{dec}}^c$ with the weighted average from the encoder, i.e. encoder context vector h^*. This concatenation is projected to the size of the vocabulary minus 1, as we do not want the model to predict the <nohs> token.

C.1.2 Copy model

The encoder part of this model is the same as that of the model explained in the previous subsection. However, this model uses various versions of the input and the decoder is altered to accommodate the copy mechanism.

First of all, we keep track of the unknown tokens in the input to provide the ability to predict them in the decoder phase. For this, we map the input utterance to temporary indices in a new extended vocabulary. This extended vocabulary contains the unknown words existing in the input utterance in

their original forms appended to the end of the original vocabulary. Since we do not want <nohs> to be predicted, we take additional precautions when it exists in the encoder input. The decoder input stays the same with unknown embeddings; nevertheless, the target utterance can include temporary indices assigned to unknown words encountered in the given input utterance, so that we can calculate the loss according to them as well.

The attention mechanism works in the same manner as in the previous model. However, we change what comes afterwards in line with the copy mechanism, where the attention for each word in the input utterance is added to their generation probabilities in the vocabulary. Here, we scatter the attention scores for the temporary indices of unknown words onto the distribution of the extended vocabulary, as well. For this reason, we maintain multiple versions of the input and output (mapped to the reduced vocabulary and mapped to the full vocabulary), as well as keeping track of the set of unknown words in the previous utterance and their temporary indices. Crucial here is the calculation of the generation probability p_{gen}, which requires the addition of several more linear layers that process the encoder context vector h_t^*, decoder input x_t, and the current decoder state s_t. As compared to the calculation of p_{gen} by See et al. (2017), we altered the formula for this value by adding $tanh$ non-linearities: $p_{gen} = \sigma(\text{tanh}(w_{h^*}^T h_t^*)+\text{tanh}(w_s^T s_t)+\text{tanh}(w_x^T x_t))$.

C.2 Reference resolution model

In this model, BERT embeddings go through a dropout layer, then a linear layer projecting the size to hidden dimensions. Finally, ReLU is applied (Nair and Hinton, 2010).

All 6 images in the context are concatenated and the concatenation goes through dropout, a linear layer and ReLU to produce the final visual context vector. We then concatenate each of the BERT representations with the visual context vector to obtain multimodal token representations. This multimodal vector goes through a linear layer and ReLU, which finalises the multimodal input vectors. The model then determines the attention to be paid to each of the multimdal vectors as indicated below:

$$e_i = v_a(tanh(W_e h_i)) \quad (6)$$

h_i is the multimodal output for each token, W_e is a linear layer projecting from hidden dimensions to attention dimensions, v_a is a linear layer that projects the output from the attention dimensions to a scalar. The model than masks the pad tokens before applying softmax over e_i scores to obtain the attention weights a_i:

$$a_i = softmax(e_i) \quad (7)$$

The final multimodally-encoded utterance representation is then the weighted average of all h_i, given their attention weights a_i:

$$h_L = \sum_i a_i h_i \quad (8)$$

Candidate images also separately go through dropout, a linear layer and ReLU. Finally, we normalise the outcomes for each image separately with L2 normalisation.

The history of each candidate image is determined by looking at their respective chains in the given game. Crucially, we only look at the chain items that were uttered before the current utterance we are trying to resolve. We take only the last utterance in the history, if such a history exists for a candidate image. In this case, we take the average of the BERT representations in the last utterance for that image. This average then goes through dropout, a linear layer and ReLU.

The final history representation for a candidate image is added to this image's final visual representation to obtain its final candidate representation. Please note that not all images in the context necessarily have histories associated with them. Therefore, some candidate representations will be multimodal, whereas the others will remain in the visual domain, with no linguistic history being added.

To determine the target image, we take the dot product between the candidate representations and the multimodally-encoded utterance representation. The candidate with highest value is then predicted to be the referent of the input utterance.

Ablation: As an ablation of the model described above, we train another type of model where the history is not added to the candidate images. Hence, the candidates are always represented only in the visual modality.

Baseline: This model only uses one-hot vectors based on image IDs. These vectors go through the same operations as the image features go through in the models described above (dropout, linear layer, ReLU and normalization). At the end, instead of a dot-product, the outputs for the candidates are

Model	Runtime
Baseline	1 hour
Proposed	5.5 hours
Ablation	2.8 hours

Table 5: Resolution: approximate training runtimes.

Model	Runtime
Ref	6.5h
ReRef	7.5h
Copy	14h

Table 6: Generation: approximate training runtimes.

Model	Parameters
Baseline	182K
Proposed	8.9M
Ablation	8.5M

Table 7: Resolution models: number of parameters.

Model	Parameters
Ref	16.1M
ReRef	24.9M
Copy	24.0M

Table 8: Generation models: number of parameters.

projected to scalar values and the model tries to predict the target via applying softmax directly over these scalars.

D Evaluation metrics

For the evaluation of the reference resolution models, we use accuracy and mean reciprocal rank (MRR) implemented by us. Accuracy is a stricter measure as it is either 0 or 1 for a given instance.

For the generation models, we use the *compute_metrics* function provided in the library at https://github.com/Maluuba/nlg-eval to obtain corpus-level BLEU, ROUGE, and CIDEr.

We also report BERTScore (Zhang et al., 2020) for the generation models. To obtain this score, we use the library provided by the authors at https://github.com/Tiiiger/bert_score and import the *score* function in our evaluation scripts. We use the BERT-uncased-model, we do not apply rescaling to baseline or importance weighting. The hash code for BERTScore that we used in evaluation is 'bert-base-uncased_L9_no-idf_version=0.3.2(hug_trans=2.6.0)'. We obtain precision, recall and F1 variants of BERTScore.

E Model configurations and reproducibility

The models are implemented in Python 3.7.5[17] and PyTorch 1.4.1[18]. In training our models, we use the Adam optimizer (Kingma and Ba, 2015) to minimize the Cross Entropy Loss with sum reduction.[19]

We experimented with learning rate (0.001, 0.0001, 0.00001), dimensions for the embeddings (512, 1024), hidden and attention dimensions (512, 1024), batch size (16, 32) and dropout probability (0.0, 0.3, 0.5). We selected the best configurations per model type via manual tuning.

We train each model type with their selected configuration with 5 different random seeds setting the random behaviour of PyTorch and NumPy. We also turn off the cuDNN benchmark and also set cuDNN to deterministic.

In all the models, the biases in linear layers were set to 0 and the weights were uniformly sampled from the range (-0.1, 0.1). In the models that learn embeddings from scratch, embedding weights were initialised uniformly in the range (-0.1, 0.1). The hidden and cell states of the LSTMs were initialised with task-related input at the first timestep.

Computing infrastructure: The models were trained and evaluated on a computer cluster with Debian Linux OS. No parallelization was implemented, each model used a single GPU GeForce 1080Ti, 11GB GDDR5X, with NVIDIA driver version 418.56 and CUDA version 10.1.

Average runtimes: Please see Table 5 and 6. These durations indicate the total approximate runtime of training. The best models are reached in a shorter amount of time.

Number of parameters in each model: Please see Table 7 and Table 8.

E.1 Configurations of the reference resolution models

We select the reference resolution models based on their performance in accurately predicting the correct target among 6 images. We also report MRR, as it also provides further information in terms of the ranking of the correct image among the distractors.

After hyperparameter search, we decided on a batch size of 32, a learning rate of 0.0001, atten-

[17] https://www.python.org/downloads/release/python-375/

[18] https://pytorch.org/

[19] Copy model in fact uses the Negative Log-Likelihood Loss that receives log-softmax probabilities. This is equivalent to Cross Entropy Loss with logits.

Model	BLEU-2	ROUGE	CIDEr	BERT-F1	ACC	MRR
Ref	22.40 (1.22)	31.29 (1.56)	41.26 (3.18)	55.24 (1.38)	59.69 (3.48)	74.41 (2.21)
ReRef	45.41 (0.89)	51.14 (0.42)	127.08 (4.17)	67.94 (0.23)	91.70 (1.09)	95.32 (0.70)
Copy	36.44 (0.31)	43.00 (0.35)	104.27 (1.16)	62.93 (0.21)	83.28 (0.77)	90.07 (0.49)

Table 9: Average metric scores of the 3 generation models on the validation set. We report the average of 5 runs and standard deviations in parentheses. ACC is the reference resolution accuracy of the sentences generated by the generation models and MRR is their mean reciprocal rank as obtained through our best reference resolution model.

tion and hidden dimensions both set to 512, and a dropout probability of 0.5 for the proposed reference resolution model. We trained the ablation model with the same settings.

Subset	ACC	MRR	Instances
First	81.85 (0.45)	88.88 (0.29)	2503
Later	88.51 (0.19)	93.33 (0.12)	3749
Overall	85.85 (0.10)	91.55 (0.07)	6252

Table 10: Validation set scores of the reference resolution model: averages of 5 runs with the best configuration, with the standard deviations in parentheses.

E.2 Configurations of the generation models

Best-performing generation models for each model type were selected based on their performance with respect to the F1 component of BERTScore. We also performed hyperparameter search for beam width used in decoding, after which we decided to use a beam width of 3. The best-performing model for each model type outperformed the other models in its own category over all metrics.

As revealed by hyperparameter search, all reported generation models use 1024 dimensions for embeddings and 512 dimensions for hidden and attention layers. They all use a learning rate of 0.0001. Ref and Copy models use a batch size of 32 and the ReRef model, 16. Ref and ReRef models use a dropout probability of 0.3, whereas the Copy model yielded better results without dropout.

F Results on the validation set

For each model we report in the main text, we also provide the validation set performances in Table 9 for the generation and Table 10 for the resolution models.

G Linguistic measures

The linguistic measures used were chosen to quantitatively explore whether artefacts of the compression, reuse and grounding present in the human utterances, as well as other human-like linguistic patterns, can be seen in the generated utterances. We compare performance of the generation models with regards to the similarity of their generated sentences to human traits, namely a) whether there is a change in token use between first and last mention (Table 11) and b) whether this relative distance, or the values in the first mention differ significantly between human and model references (Table 12).

In the case of givenness markers, we measure this as the proportion of tokens which correspond to definite (*the*), indefinite (*some, a, an*) and other markers of the existence of shared context (*again, before, one, same, also*) which occur in the utterance. In the case of compression, we measure the lengths of the utterances in terms of tokens, and content tokens (tokens which are not in the stopword list from from nltk version 3.4.5 (Loper and Bird, 2002). We also measure the proportion of content words in an utterance which correspond to nouns, verbs and adjectives. Finally, for entrainment, examining only later utterances (not the first referent to an image), we measure firstly what proportion of the utterance in question consists of reused unigrams and bigrams from the previous utterance. We also measure within the reused tokens, the proportion of which is made up of nouns, adjectives and verbs, in order to discover their relative importance in terms of reuse. These measures can all be found in Tables 11 and 12. For these analyses we compared the generated output from the best seed for each model variant. These were seeds 1, 1, and 24 for the Ref, Copy and ReRef models respectively. We report both effect size (d) as measured by Cohen's d, and p-value ($*p < 0.05$, $**p < 0.005$, $***p < 0.001$) for each comparison. We use the Scipy stats package (scipy version 1.3.3.) *ttest_ind* to perform the independent t-test, and our own implementation to calculate Cohen's d effect size.

Additionally to check general fluency, we evaluate the coherence and vocabulary use of the models in comparison to humans. We measure *Type Token Ratio (TTR)*, the proportion of unique tokens

	Human			*ReRef*			*Copy*			*Ref*		
	first	later	d	first	later	d	first	later	d	first	later	d
Givenness												
givenness	0.05	0.08	-0.36*	0.02	0.10	-0.89*	0.04	0.09	-0.53*	0.05	0.05	-0.03
definite	0.03	0.05	-0.27*	0.01	0.08	-0.85*	0.03	0.06	-0.48*	0.04	0.05	-0.04
seen	0.01	0.03	-0.26*	0.00	0.02	-0.43*	0.01	0.03	-0.29*	0.00	0.00	0.03
indefinite	0.07	0.02	0.77*	0.15	0.01	1.88*	0.10	0.01	1.14*	0.15	0.15	0.03
Compression												
length_c	11.29	8.28	0.63*	11.32	7.22	1.15*	10.77	7.79	0.65*	13.66	13.59	0.00
prop content	0.53	0.57	-0.20*	0.41	0.54	-0.70*	0.50	0.58	-0.39*	0.40	0.39	0.01
prop noun	0.37	0.41	-0.29*	0.30	0.44	-0.86*	0.37	0.43	-0.37*	0.28	0.28	-0.01
prop adj	0.09	0.10	-0.02	0.06	0.07	-0.14*	0.08	0.09	-0.10*	0.08	0.08	0.04
prop verb	0.13	0.11	0.12*	0.19	0.11	0.76*	0.13	0.12	0.12*	0.17	0.17	0.01

Table 11: Trends in Subsequent mentions across humans, ReRef, Copy and Ref. The presence of * indicates significant differences between first and later means, with $p < 0.001$. d shows effect size measured by Cohen's d.

Human		*ReRef*			*Copy*			*Ref*		
	mean	mean	d	p	mean	d	p	mean	d	p
Lexical Entrainment:										
reuse prop within mention:										
–reuse_c	0.562	0.660	-0.334	***	0.612	-0.168	***	0.320	0.868	***
–reuse_bigrams_c	0.325	0.304	0.050	*	0.283	0.103	***	0.091	0.682	***
reuse prop within reused:										
–noun	0.701	0.746	-0.161	***	0.716	-0.050	*	0.740	-0.124	***
–adj	0.158	0.146	0.054	*	0.146	0.057	*	0.180	-0.079	**
–verb	0.095	0.066	0.165	***	0.097	-0.011	0.653	0.063	0.172	***
–NN bigrams	0.064	0.051	0.069	**	0.056	0.043	0.064	0.013	0.328	***

Table 12: Human comparison with ReRef, Copy and Ref for givenness markers and Compression. The presence of * indicates a significant difference between the human mean and that of the model. *(***: $p < 0.001$, **: $p < 0.005$, *: $p < 0.01$)*

in an utterance. This can capture ungrammatical repetition patterns in the generation, and, if following human trends, should increase in subsequent mentions. Although both models have significantly lower TTR than the human data, ReRef, unlike Copy, shows a significant increase in subsequent mentions, with much higher TTR than Copy, even though both models show similar average utterance length for later utterances *(ReRef: 7.22, Copy: 7.79)*. In terms of vocabulary, for the generated outputs, ReRef has a much smaller *(first: 492, later: 705)* vocabulary than Copy *(first: 1098, later: 1469)*, although these are both much lower than Human vocabulary size *(first: 1836, later: 1727)* and show an increase rather than a decrease in later mentions.

Overall, Tables 11 and 12 show that both of our context-aware speaker models ReRef and Copy are able to generate referring utterances which make use of the dialogue history in a manner akin to humans with respect to multiple aspects of language style.

Comparing the context-aware models, ReRef shows a stronger degree of shortening than Copy, with very similar levels of bigram reuse to humans while Copy shows more similar traits to humans in terms of proportion of markers and PoS tags (as revealed by smaller effect sizes). In general, both models are successful at generating human-like utterances as we measure them, however it seems that while Copy does generate utterances with the most similar proportional similarities to humans and exhibits similar proportions of unigram reuse, it does so at the expense of coherence. In terms of content bigram reuse, Copy seems to be less selective in what it repeats from previous referring utterances than ReRef, most likely due to the increased overall level of repetition in the generation. ReRef on the other hand shows amplified versions of the human trends, yet very similar content bigram and noun-noun bigram reuse proportion to humans, while maintaining low levels of same content word repetition as well as a high TTR, which indicates that coherence is also maintained.