DialEdit: Annotations for Spoken Conversational Image Editing

Ramesh Manuvinakurike[2*], **Jacqueline Brixey**[2*], **Trung Bui**[1],
Walter Chang[1], **Ron Artstein**[2], **Kallirroi Georgila**[2]
[1]Adobe Research
[2]Institute for Creative Technologies, University of Southern California
[manuvinakurike,brixey,artstein,kgeorgila]@ict.usc.edu
[bui,wachang]@adobe.com

Abstract

We present a spoken dialogue corpus and annotation scheme for conversational image editing, where people edit an image interactively through spoken language instructions. Our corpus contains spoken conversations between two human participants: users requesting changes to images and experts performing these modifications in real time. Our annotation scheme consists of 26 dialogue act labels covering instructions, requests, and feedback, together with actions and entities for the content of the edit requests. The corpus supports research and development in areas such as incremental intent recognition, visual reference resolution, image-grounded dialogue modeling, dialogue state tracking, and user modeling.

1 Introduction

Photographs have emerged as a means for sharing information, effective storytelling, preserving memories, and brand marketing among many other applications. The advent of photo-centric social media platforms such as Instagram, Snapchat, etc. along with easy access to high quality photo-taking devices has only made photographs a more powerful medium.

Photographs are often edited with the intention of improving their quality (e.g., fixing the lighting), for use in a narrative (e.g., for an ad campaign), for alteration (e.g., removing objects from the image), for preservation of memories (by restoring old photographs), and for other reasons. Social media platforms support popular and extensively used editing methods called presets (or filters). Such presets can also be found in cameras on many current smartphones, and can be applied to photographs almost instantaneously. However, image editing is far from choosing the right filter or preset values. Photo editing is a complex task often involving diligently and skillfully executed steps that require expertise.

Seeking professional help for editing photographs is common, and can be seen in popular forums such as Reddit Photoshop Request (https://www.reddit.com/r/PhotoshopRequest/) and Zhopped (http://zhopped.com/), where users post their photographs and request help from professionals. The professionals then either volunteer for free or do the job for a fee. The process typically starts with users publicly posting their request and the photograph they desire to be edited. These requests are formulated in an abstract manner using natural language (Ex: "I love this photo from our trip to Rome. Can someone please remove my ex from this photo? I am the one on the right."), rather than intricate multi-step instructions (Ex: "Free select the person on the left, replace the region with the building on the bottom left using replacement tools, fix the blotch by the healing tool..."). The professionals download these photographs, edit them, and post them back. They have knowledge about the image editing tool used, skills, time, and artistic creativity to perform the changes. If the users are not happy with the results, they post their modification requests, and then the professionals incorporate these changes and post the updated photographs. While these forums are popular, such methods have a few drawbacks. Because the expert editors edit the photographs without the requester being able to see the changes being performed in real time, (i) the users are not able to provide real-time feedback; (ii) it is hard for the users to provide requests for all needed modifications; and (iii) the professional editors cannot ask for minor clarifications while editing the photographs. These drawbacks often result in modifications that do not match the users'

[*] Work done while at Adobe Research.

expectations. The alternative solution of the users performing the edits themselves is difficult and time consuming as the image editing tools have a steep learning curve.

Our ultimate goal is to develop a conversational agent that can understand the user requests, perform the edits, guide the user by providing suggestions, and respond in real time. In this paper we present a novel corpus that captures the conversation between the user who wants to edit a photograph and the expert human wizard who performs the edits (playing the role of a future dialogue system). We introduce a novel annotation scheme for this task, and discuss challenging sub-tasks in this domain. Conversational image editing combines spoken language, dialogue, and computer vision, and our real-world domain extends the literature on domains that are at the intersection of language and computer vision. We will publicly release our corpus in the near future.

2 Related Work

Conversation in the context of visual information has been studied for a long time. Clark and Wilkes-Gibbs (1986) studied reference resolution of simple figures called tangrams. Kennington and Schlangen (2015) and Manuvinakurike et al. (2016) performed incremental understanding and incremental reference resolution respectively in a domain of geometric shape descriptions, while Schlangen et al. (2016) resolved references to objects in real-world example images. Much work has been done in the context of gamified scenarios where the interlocutors interact and resolve references to real-world objects (Kazemzadeh et al., 2014; Paetzel et al., 2014; Manuvinakurike and DeVault, 2015). Also, such gamified scenarios have served as platforms for developing/learning incremental dialogue policies regarding whether the system should respond immediately or wait for more information (Paetzel et al., 2015; Manuvinakurike et al., 2017). Referential domains in the context of dialogue have also been studied using virtual reality technologies and spatial constraints (Stoia et al., 2008; Das et al., 2018) as well as robots (Whitney et al., 2016; Skantze, 2017).

A more recent direction of research involving dialogue and vision has been in the context of answering factual questions on images (Das et al., 2017; Antol et al., 2015) using the MSCOCO data set (Lin et al., 2014). The task may also involve a gamified scenario with the interlocutors playing a yes-no question-answer game as in de Vries et al. (2017). In these works the focus is less on the dialogue aspects and more on the factual aspects of the images, i.e., if an object is present or what a certain component of the image is. Mostafazadeh et al. (2017) extended this line of work with conversations grounded on images. Furthermore, Huang et al. (2016) built a data set of images with corresponding descriptions in sequence, for the task of visual storytelling.

Other gamified real-world scenarios involve object arrangement (DeVault and Stone, 2009), puzzle completion (Iida et al., 2010; Takenobu et al., 2012), map navigation (Anderson et al., 1991; Lemon et al., 2001; Johnston et al., 2002), furniture-buying scenarios (Di Eugenio et al., 2000), and treasure-hunt tasks in a virtual environment (Byron and Fosler-Lussier, 2006). A multi-modal interface for image editing combining speech and direct manipulation was developed by (Laput et al., 2013). With this interface a user can for example select a person's hat in an image and say "this is a hat". Then the system learns to associate the tag "hat" with the selected region of the image. Finally, Manuvinakurike et al. (2018a) recently introduced a corpus containing one-shot image editing instructions.

3 Data

The task of image editing is challenging for the following reasons: (i) The user needs to understand whether changes applied to a given image fit the target narrative or not. (ii) Image editing is a time consuming task. The user typically experiments with various features often undoing, redoing, altering in increments, or even completely removing previously performed edits before settling on the final image edit. (iii) The users may know at an abstract level what changes they want to perform, but be unaware of the image editing steps or parameters that would produce the desired outcome. (iv) Image editing tools are complicated due to the availability of innumerable options, and can have a steep learning curve often requiring months of training.

Our task is particularly well suited for spoken dialogue research. Besides understanding the user utterances and mapping them to commands supported by the tool, the task also involves a high degree of interactivity that requires real-time understanding and execution of the user requests. For instance, in a dialogue setting and in order to increase the saturation value, the user can utter "more, more, more" until the desired target value has been set. An annotation scheme should support such incremental changes as well as requests for new changes, updates of ongoing changes (including undoing and redoing), comparing the current version of the image with previous versions, and question-answer exchanges between the user and the wizard (including suggestions, clarifications, and feedback).

3.1 Data Collection

We collected spoken dialogues between users (who request image edits) and wizards (who perform the edits); a total of 28 users and 2 wizards participated in the collection. Prior to data collection, our wizards (the first two authors) were trained in executing a range of image edits.

We tested several image editing tools and found that very simple tools that did not support a high degree of functionality resulted in extremely restrictive dialogues lacking variety. Conversely, tools with rich functionality, such as Adobe Photoshop or GNU GIMP, resulted in user image edit requests that required hours to complete. Such interactions yielded creative image edit requests but did not yield timely dialogue phenomena. The tool ultimately used for image editing in this study was Adobe Lightroom. This tool produced diverse and highly interactive dialogues for image editing. The tool is popular among photographers and supports a wide variety of functionality. Users were able to make creative requests with few restrictions, and these requests could often be executed rapidly.

3.2 Experiment Setup

The recruited users were given images (digital photographs) sampled from the Visual Genome data set (Krishna et al., 2017) which in turn were sampled from the MSCOCO data set (Lin et al., 2014). The photos selected from the sampled image data sets were based on observations of 200 random request submissions from Zhopped and Reddit Photoshop forums. The forum submissions were often about eight high-level categories of images: animals, city scenes, food, nature/landscapes, indoor scenes, people, sports, and vehicles. Thus we selected images from the MSCOCO data set that fit into at least one of these eight categories.

Users were given one photograph from each category in an experiment session. They were given time to think about the changes they wanted to perform before the dialogue session, and were informed about the tool that was going to be used and the fact that it did not support complex functionality. If they were unsure of what functionality was supported they were instructed to ask the wizard. Users were asked to perform as many edits as they desired per image. Participants were encouraged (but not required) to participate for 40 minutes, and communicated via remote voice call. Users did not have the freedom to perform the edits themselves. Any edits they wished to be performed on the image had to be conveyed to the wizard through voice. The wizard responded to the requests in a natural human-like manner. The screen share feature was enabled on the wizard's screen so that the user could see in real time the wizard's edits on the image. While users were not explicitly told that the wizard was human, this was obvious due to the naturalness of the conversation.

The interaction typically started with the user describing a given image to the wizard. The wizard was not aware of the images provided to the user. The wizard chose the image from the available images based on the user description; following user confirmation, the image was then loaded for editing. The image editing session generally began with the user describing desired changes to the image in natural language. The wizard interpreted the request provided by the user and performed these edits on the image. The interaction continued until the user was satisfied with the final outcome. Figure 1 shows an example of an interaction between the user and the wizard.

3.3 Annotation Scheme

We designed a set of 26 dialogue act types, for the ultimate goal of building a conversational agent. Some of the dialogue acts were motivated by Bunt et al. (2012), while others are specific to the domain

Figure 1: Sample interaction between the user and the wizard.

Dialogue Act	**Description**
Image Edit Request (IER)	user requests changes to the image (IER-N, IER-U, IER-R, IER-C)
Comment (COM)	user comments on the image or edits (COM-L, COM-D, COM-I)
Request Recommendation (RQR)	user requests recommendation from the wizard on editing ideas
Question Feature (QF)	user asks question on the functionality of the editing tool
Question Image Attribute (QIA)	user asks question about the image
Request Feedback (RF)	user requests feedback about the image edits
Image Location (IL)	user & wizard locate the image at the beginning
Action Directive (AD)	user asks wizard to act on the application, e.g., "click the button"
Finish (FIN)	user wants to end the editing session
Suggestions (S)	wizard suggests ideas for editing the image
Request IER (RQIER)	wizard requests user to provide IER
Confirm Edit (CE)	wizard confirms the edit being performed
Feature Preference (FP)	wizard requests which tool option to use for achieving the user edits
Narrate (N)	wizard gives narration of the steps being performed
Elucidate (E)	these are wizard responses to QF & QIA
No Support (NS)	wizard informs user that the edit is not supported by the tool
Respond Yes/No (RSY/RSN)	yes/no response
Acknowledge (ACK)	acknowledgment
Discourse Marker (DM)	discourse marker
Other (O)	all other cases

Table 1: Dialogue act types.

of conversational image editing. Dialogue acts apply to segmented utterances, with each segment annotated with one dialogue act. Note that an utterance is defined as a portion of speech preceded and/or followed by a silence interval greater than 300 msec. Most of the dialogue act types are summarized in Table 1; below we elaborate on three specific classes: image edit requests (IER), comments (COM), and suggestions (S).

Image Edit Requests (IER): Image edit requests are grouped into four categories. New requests (IER-N) are edits that the users desire to see in the image, which are different from previous requests. Update requests (IER-U) are refinements to a previous request (users often request updates until the target is achieved). Revert requests (IER-R) occur when users want to undo the changes done to the image until a certain point. Compare requests (IER-C) occur when users want to compare the current version of the image to a previous version (before the most recent changes took place). The image edit requests IER-N and IER-U are labeled further with action and entity labels, which specify the nature of the edit request (the use of actions and entities is inspired by the intents and entities of Williams et al. (2015)). These labels serve as an intermediary language to map a user's utterance to executable commands that can be carried out in an image editing program. Actions are a predefined list of 18 functions common to most

Segments	Dialogue Act	Action	Attribute	Loc/Obj	Mod/Val
uh	O	-	-	-	-
make the tree brighter	IER-N	Adjust	brightness	tree	-
like a 100	IER-U	Adjust	brightness	tree	100
nope too much	COM-D	-	-	-	-
perfect	COM-L	-	-	-	-
let's work on sharpness	IER-N	Adjust	sharpness	-	-

Table 2: Example annotations of dialogue acts, actions, and entities.

Dialogue Act	% Words	% Utterance Segments	Dialogue Act	% Words	% Utterance Segments
IER-N	19.4	9.2	FIN	1.5	1.0
IER-U	16.3	12.5	S	4.7	4.0
IER-R	1.0	0.8	RQIER	2.1	2.6
IER-C	0.5	0.3	CE	1.6	1.9
COM-L	4.9	6.0	FP	0.1	0.1
COM-D	1.8	1.5	N	3.1	4.2
COM-I	2.5	1.5	E	1.3	0.7
RQR	0.7	0.0	NS	1.0	0.6
QF	1.1	0.6	RSY	2.3	6.8
QIA	0.3	0.2	RSN	0.9	1.2
RF	0.0	0.0	ACK	6.4	17.6
IL	3.0	1.5	DM	2.3	6.5
AD	4.8	3.9	O	16.4	14.8

Table 3: Percentages of words and of utterance segments for each dialogue act type; "0.0" values are close to 0.

image editing programs, such as cropping. Each IER contains at most one action. The entities provide additional information without which the action cannot be applied to the given image. The entities are made up of attributes (saturation, contrast, etc.), region/object (location where the image edit action is to be applied), value (modifiers or cardinal values accompanying the action-attribute). Table 2 shows example annotations.

Comments (COM): Three types of user comments are annotated: (i) Like comments (COM-L) where users show a positive attitude towards the edits that are being performed ("that looks interesting", "that's cool"). (ii) Dislike comments (COM-D) are the opposite of like comments ("I don't like that", "I don't think it's what I want"). (iii) Image comments (COM-I) are neutral user comments such as comments on the image ("it looks like a painting now", "her hair looks pretty striking").

Suggestions (S): Suggestions were the recommendations issued by the wizards to the users recommending the image editing actions. Suggestions also included the utterances that were issued with the goal of helping the user achieve the final image edits desired.

Table 3 shows the prevalence of our 26 dialogue acts in the corpus (percentage of words and of utterance segments in the corpus per dialogue act).

3.4 Data Preparation

The conversations were recorded using the OBS software which is a free open-source program for streaming video and audio. Then the audio data were extracted from the videos. Transcription was done on small audio chunks which was more convenient and faster than transcribing long clips. The

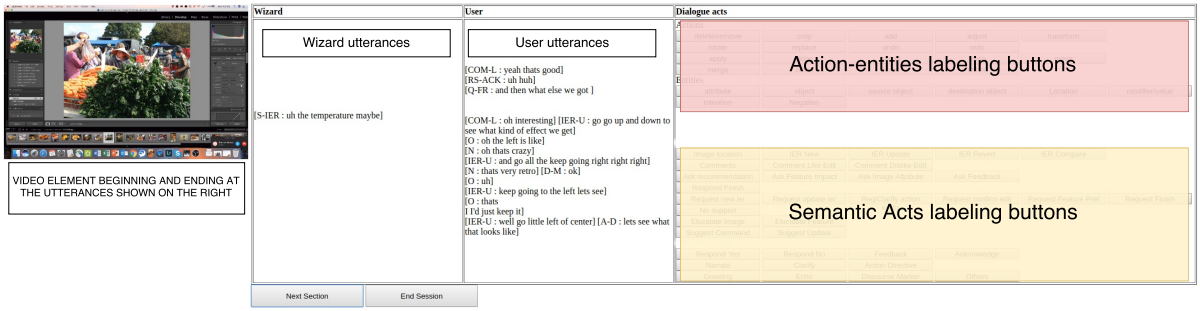

Figure 2: Web annotation tool used to annotate the dialogue. The figure shows the wizard and the user utterance aligned with time. The video element is shown to the left. The annotation is performed by highlighting the text and clicking the buttons corresponding to the dialogue act.

small audio clips were obtained by splitting the audio at the silence points using the webRTC Voice Activity Detection (https://pypi.python.org/pypi/webrtcvad). Transcriptions were performed using the Amazon MTurk platform. The transcribed audio data were then validated and annotated with dialogue acts, actions, and entities using a custom-built web annotation tool (Figure 2). The annotations were performed by two expert annotators who were well versed with the annotation scheme. Figure 2 shows the tool that was built for annotating the dataset. The tool was web-based, with the annotators being able to see the video, audio interaction and the transcriptions shown in small chunks (typically around 45 seconds) which were to be annotated by selecting the text and the corresponding dialogue act. In order to calculate the validity of the annotation scheme we calculated inter-rater reliability for dialogue act labeling by having two expert annotators annotate a single dialogue session; kappa was 0.81. In total 28 users contributed to 129 dialogues with 8890 user utterances, 4795 wizard utterances, and 858 minutes of speech. The total number of tokens in the user and wizard utterances is 59653 and 26284 respectively. Also, there are 2299 unique user tokens, 1310 unique wizard tokens, and 2650 total unique tokens.

4 Discussion

The transitions between dialogue acts for the user acts were analyzed (for this analysis we ignore the label "Other-O"). We found that the most common transition was from IER-U to IER-U. This is particularly interesting as it shows that users provide a series of updates before attaining the final image edits. This transition was more common than IER-N to IER-U, which is the second most frequently found transition. Users were found to like the image edits after IER-Us, and after issuing a COM-L (like edit) comment they usually move on to the next IER. We also found that when users disliked the edits (COM-D) they did not entirely cancel the edits but continued updating (IER-U) their requests until the final image version fit their needs. Transitions from IER-N to IER-N were also common; users could issue a complete new image edit request IER-N and then move on to another new image edit request IER-N.

The corpus can support research on the following (but not limited to) challenging sub-tasks:

Object detection: Understanding to which objects or regions the user refers in the image edit requests needs object identification. This is an actively researched topic in the computer vision research community.

Dialogue act labeling: Human speech is spontaneous, ungrammatical, and filled with disfluencies, among many other characteristics. Understanding the user intentions through dialogue act labeling on spontaneous human speech is a challenging problem. Our corpus has similar challenges as the Switchboard data set (Godfrey et al., 1992), however, in our case the dialogue takes place in a situated domain involving a visual environment. Our corpus has recently been used for incremental dialogue act identification (Manuvinakurike et al., 2018b).

State tracking: Dialogue state tracking means accurately tracking the user goals during a dialogue. In our work, state tracking refers to tracking the users' goals as they are making edits to an image.

Dialogue management: Designing a dialogue policy for this task is challenging due to the instan-

taneous and rapid nature of the interaction. A good dialogue policy should support incrementality. For example, users would often say "more, more, more" until the desired value of saturation was obtained. Thus the dialogue system should be able to process the user's utterance and perform the corresponding actions, as soon as the user's speech becomes available (Manuvinakurike et al., 2018b). Incrementality in dialogue is analogous to input autocomplete or typeahead used in search engines which accelerates the user's interaction by predicting the full query intention as a user is typing. Furthermore, the dialogue manager should be capable of generating the right utterances so that the interaction results in the desired image.

Nature of interactions: The task is very interactive. The users provide feedback and issue image edit updates in real time, which means that the user's input needs to be tracked in real time. The like (COM-L) and dislike (COM-D) user comments can be useful for tracking the likelihood that the user will keep the edits. The wizards are usually not static performers but also need to track the changes occurring in the image, and play an important role in helping the users achieve their goal. The wizards issue suggestions to the users when they need help with editing the images and issue clarifications about the tool and features supported by it (e.g., "User: Can we fade the picture? Wizard: We can try the clarity tool.").

5 Conclusion

We presented a novel spoken dialogue corpus on "conversational image editing". We described our data collection process and novel dialogue act labeling scheme. Our annotation scheme consists of 26 dialogue act labels covering instructions, requests, and feedback. The corpus supports research and development in areas such as incremental intent recognition (Manuvinakurike et al., 2018b), dialogue modeling, and dialogue state tracking. Furthermore, the data set is constructed using richly annotated images, which makes it an ideal platform for studying reference resolution in images, question answering, image-grounded dialogue modeling, tracking user likeness of images, and user modeling (providing suggestions to users depending on their preferences and knowledge of the tool). The corpus will be publicly released in the near future.

Acknowledgments

Apart from his internship at Adobe, the first author was also supported by a generous gift of Adobe Systems Incorporated to USC/ICT. The second author thanks the National GEM Consortium for fellowship funding and internship placement at Adobe. The last two authors were supported by the U.S. Army; statements and opinions expressed do not necessarily reflect the position or policy of the U.S. Government, and no official endorsement should be inferred.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of ICCV*, pages 2425–2433, Santiago, Chile.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R. Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of LREC*, pages 430–437, Istanbul, Turkey.

Donna K. Byron and Eric Fosler-Lussier. 2006. The OSU Quake 2004 corpus of two-party situated problem-solving dialogs. In *Proceedings of LREC*, pages 395–400, Genoa, Italy.

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

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 CVPR*, pages 326–335, Honolulu, Hawaii, USA.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of CVPR*, Salt Lake City, Utah, USA.

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 CVPR*, pages 5503–5512, Honolulu, Hawaii, USA.

David DeVault and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of EACL*, pages 184–192, Athens, Greece.

Barbara Di Eugenio, Pamela W. Jordan, Richmond H. Thomason, and Johanna D. Moore. 2000. The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, 53(6):1017–1076.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, San Francisco, California, USA.

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of NAACL–HLT*, pages 1233–1239, San Diego, California, USA.

Ryu Iida, Shumpei Kobayashi, and Takenobu Tokunaga. 2010. Incorporating extra-linguistic information into reference resolution in collaborative task dialogue. In *Proceedings of ACL*, pages 1259–1267, Uppsala, Sweden.

Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. 2002. MATCH: An architecture for multimodal dialogue systems. In *Proceedings of ACL*, pages 376–383, Philadelphia, Pennsylvania, USA.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of EMNLP*, pages 787–798, Doha, Qatar.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of ACL*, pages 292–301, Beijing, China.

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.

Gierad P Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. PixelTone: A multimodal interface for image editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2185–2194, Paris, France.

Oliver Lemon, Anne Bracy, Alexander Gruenstein, and Stanley Peters. 2001. Information states in a multi-modal dialogue system for human-robot conversation. In *Proceedings of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue (Bi-Dialog)*, pages 57–67.

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 *Proceedings of the European Conference on Computer Vision*, pages 740–755, Zurich, Switzerland.

Ramesh Manuvinakurike and David DeVault. 2015. Pair me up: A web framework for crowd-sourced spoken dialogue collection. In Gary Geunbae Lee, Hong Kook Kim, Minwoo Jeong, and Ji-Hwan Kim, editors, *Natural Language Dialog Systems and Intelligent Assistants*, chapter 18, pages 189–201. Springer.

Ramesh Manuvinakurike, Casey Kennington, David DeVault, and David Schlangen. 2016. Real-time understanding of complex discriminative scene descriptions. In *Proceedings of SIGDIAL*, pages 232–241, Los Angeles, California, USA.

Ramesh Manuvinakurike, David DeVault, and Kallirroi Georgila. 2017. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of SIGDIAL*, pages 331–341, Saarbrücken, Germany.

Ramesh Manuvinakurike, Jacqueline Brixey, Trung Bui, Walter Chang, Kim Doo Soon, Ron Artstein, and Kallirroi Georgila. 2018a. Edit me: A corpus and a framework for understanding natural language image editing. In *Proceedings of LREC*, pages 4322–4326, Miyazaki, Japan.

Ramesh Manuvinakurike, Trung Bui, Walter Chang, and Kallirroi Georgila. 2018b. Conversational image editing: Incremental intent identification in a new dialogue task. In *Proceedings of SIGDIAL*, pages 284–295, Melbourne, Australia.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of IJCNLP*, pages 462–472, Taipei, Taiwan.

Maike Paetzel, David Nicolas Racca, and David DeVault. 2014. A multimodal corpus of rapid dialogue games. In *Proceedings of LREC*, pages 4189–4195, Reykjavik, Iceland.

Maike Paetzel, Ramesh Manuvinakurike, and David DeVault. 2015. So, which one is it? The effect of alternative incremental architectures in a high-performance game-playing agent. In *Proceedings of SIGDIAL*, pages 77–86, Prague, Czech Republic.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of ACL*, pages 1213–1223, Berlin, Germany.

Gabriel Skantze. 2017. Predicting and regulating participation equality in human-robot conversations: Effects of age and gender. In *Proceedings of HRI*, pages 196–204, Vienna, Austria.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: A situated corpus with annotated referring expressions. In *Proceedings of LREC*, pages 650–653, Marrakech, Morocco.

Tokunaga Takenobu, Iida Ryu, Terai Asuka, and Kuriyama Naoko. 2012. The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of LREC*, pages 422–429, Istanbul, Turkey.

David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. 2016. Interpreting multimodal referring expressions in real time. In *Proceedings of ICRA*, pages 3331–3338, Stockholm, Sweden.

Jason D. Williams, Eslam Kamal, Mokhtar Ashour, Hani Amr, Jessica Miller, and Geoffrey Zweig. 2015. Fast and easy language understanding for dialog systems with Microsoft Language Understanding Intelligent Service (LUIS). In *Proceedings of SIGDIAL*, pages 159–161, Prague, Czech Republic.