SCARE: A Situated Corpus with Annotated Referring Expressions

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron and Eric Fosler-Lussier

The Ohio State University, Computer Science and Engineering
2015 Neil Ave., Columbus, Ohio 43210
stoia—shockley—dbyron—fosler@cse.ohio-state.edu

Abstract

In this paper we report on the release of a corpus of English spontaneous instruction giving situated dialogs. The corpus was collected using the Quake environment, a first-person virtual reality game, and consists of pairs of participants completing a direction giver-direction follower scenario. The corpus contains the collected audio and video, as well as word-aligned transcriptions and the positional/gaze information of the player. Referring expressions in the corpus are annotated with the IDs of their virtual world referents.

1. Introduction

With the fast development of mobile technologies, there is an increasing interest in creating embodied conversational partners and using language interfaces for a variety of situated tasks (tasks performed from a location within an environment). Many research projects deal with situated language, from many perspectives: interpretation of situated language (Lauria et al., 2001), using visual information in the referring process (Kelleher et al., 2005), helping users navigate using hand-held tourist information portals (Johnston et al., 2002), giving pedestrian directions (Yang et al., 1999) or in-car driving direction systems (Dale et al., 2003), *inter alia*. All of these applications present an exciting and challenging new frontier for dialog agents, since attributes of the real-world setting must be combined with other contextual factors for the agent to communicate successfully.

Even though a wealth of speech data is available for the dialog systems research community, the particular field of situated language has yet to find an appropriate free resource. The corpus required to answer research questions related to situated language should connect world information to the human language. One situated language corpus is available (Byron and Fosler-Lussier, 2006), but it does not include information to automatically link world attributes with the language. In the current work, we use the same stimuli and tasks, however the roles of the dialog partners have been modified. Only one partner was placed in the world, and the other partner was given full knowledge of the world to be able to plan how to complete the tasks. This produced a large number of referring expressions and instructional language. Compared to the corpus presented in (Byron and Fosler-Lussier, 2006) this corpus contains referent annotation, synchronized positional/gaze information and word aligned transcripts. This paper serves to announce this corpus as a public resource and describe its components.

2. Data Collection Procedure

The interaction captured in the corpus takes place in a virtual reality (VR) world rendered and tracked by a game engine[1]. The game log records the user's position and orientation in the virtual world, the locations of objects and the timing of events that take place in the world.

The VR world was chosen instead of a real-world setting so that complex spatially-extended tasks could be studied without the expense of specialized equipment to obtain detailed information on context variables such as locations and view angles. The VR materials can be reused by other research groups and could be modified to suit different research questions. VR worlds are also suitable for distributed web-based data collections.

Although we used a virtual world, humans have been found to be very robust in treating virtual world spatial representations in the same way as real-world objects, even when the graphical depiction in the virtual world is very impoverished (Peruch et al., 2000). We take the view that spatial language and references to objects in a virtual world maintain most properties when transferred to the real world domain, and a corpus collected in a virtual environment will prove useful for studying general language behavior.

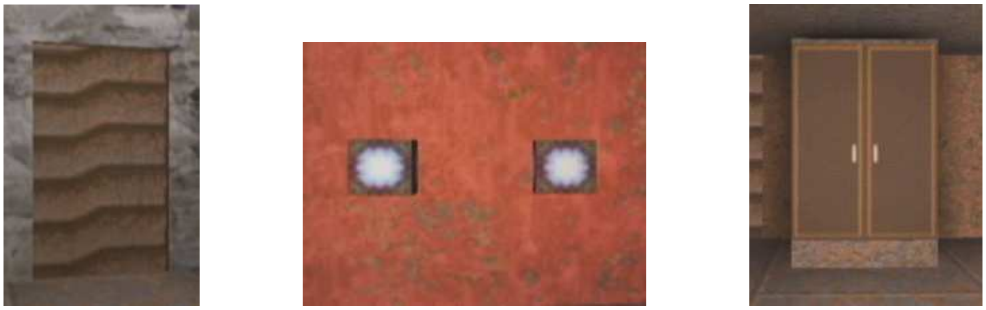
2.1. Physical configuration and tasks

This study was designed to elicit natural, spontaneous situated language examples from human partners. In the VR world, one partner, the Direction Follower (DF), moves about to perform a series of simple manipulation tasks. The simulated world was presented from a first-person perspective on the DF's desktop computer monitor. The world is a two level maze, with a total of eighteen rooms, two flights of stairs and a long hallway. The world contains only a small number of object types: buttons, cabinets, doors, tables, and so on. Figure 1 shows examples of objects that populate the VR world.

The DF had no prior knowledge of the world map or tasks and relied on his partner, the Direction Giver (DG) to guide him on completing the tasks. The DG had a paper 2D map of the world and a list of tasks to complete (such as finding treasures and hiding them in different cabinets or rearranging objects). The partners spoke to each other through headset microphones.

As the participants collaborated on the tasks, the DG had instant feedback of the DF's location in the VR world, because the game engine displayed the DF's first person view of the world on both the DG's and DF's computer monitors. Figure 2 shows an example view of the world, the map representation of the current room and the accompanying dialog fragment. The referring expressions that identify buttons, doors and cabinets are indicated in bold. Note

[1] http://www.idsoftware.com/games/quake/quake2/



(a) Door (b) Buttons (c) Cabinet

Figure 1: Example objects from the experiment world

that this corpus is not suitable for studying gestures, as the partners see a first person view of the world, and not their own bodies.

As the DF performed the task, the video stream showing his view of the world was captured to a camera, along with the audio from both microphones. A log-file created by the VR software recorded the DF's coordinates, gaze angle, and the position of objects in the world at a frequency of 10 times per second. These two data sources were synchronized using calibration markers. A technical report is available that describes the recording equipment and software used (Byron, 2005).

It is important to note that the knowledge shared by the dialog partners in this domain comes from both the dialog they are engaged in, and also their shared view of the world. The DF's actions change the state of the world, and his partner is aware of these changes through the visual input.

2.2. Demographics

All the participants identified themselves as native speakers of North American English, with an average age of 30. The participants were recruited in pairs and usually they were friends, colleagues or members of the same family. There were 19 male and 11 female participants. At the end of each data collection session, the participants completed a survey. All the questions used a 1(low/easy) to 5(high/difficult) rating scale. The participants rated themselves high on computer expertise (with an average of 4.43) and they found that the tasks were not very difficult (1.87 average), the virtual world was not hard to navigate in (1.71 average), and the descriptions they heard were not difficult to follow (1.45 average).

3. Data Preparation: Transcriptions and Annotations

Using the above-described setup, we created a corpus of 15 dialogs containing a total of 3 hours and 41 minutes of speech. The corpus was transcribed and word-aligned using Praat (Boersma and Weenink, 2001). SONIC (Pellom and Hacioglu, 2001) speech recognition software was used to automatically word align the utterances, which were corrected by two human annotators. The dialogs were further annotated using the Anvil software (Kipp, 2004), a free video annotation tool, to identify a set of target referring expressions in the corpus.

Det			**Head**		
Value	Count	Percent	Value	Count	Percent
the	364	39%	common noun	558	60%
that/this	264	29%	one	166	18%
none	253	27%	it	116	13%
a	46	5%	that	57	6%
			none	30	3%

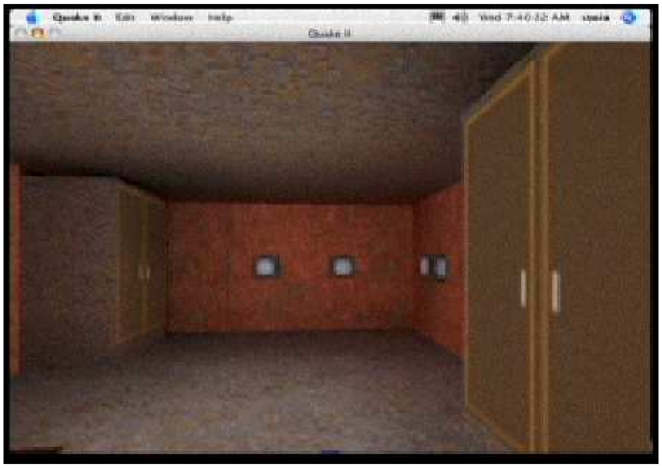
Table 1: Distribution of **Det** and **Head** values in the labeled references

4. Referring Expressions: Corpus Distribution

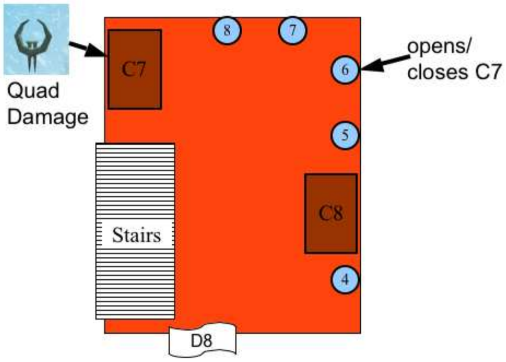
The current corpus annotation was restricted to doors, buttons, and cabinets, which are key objects in completing the tasks. The world was design to contain many instances of these objects. Referring expressions (REs) in the corpus are annotated with the appropriate symbolic identifier (ID) assigned to the object in the VR world. In the case of locative REs (e.g. *the button on the left*), the entire noun phrase, including the locative prepositional phrase (e.g. *on the left*), is included as part of the RE. Items that did not contain a surface realization of the head of the NP (e.g. *on the left*), are marked with the tag **Empty**, but the ID is still included. There are no embedded annotations (e.g. *the button next to the door* is annotated as one RE which refers to a button). Indefinite REs are annotated with IDs in the cases where it was possible for the annotators to determine the ID from the context of the task. REs with plural referents are marked as **Set**, and are labeled with a list of the members in the set. REs are also annotated as either **Vague** when the referent was not clear to the annotator at the time of utterance or **Abandoned** if the utterance was cut short (IDs are added when possible).

The corpus contains a consensus version on which both annotators (the two lead authors) agreed on the associated IDs and the properties of the REs. Due to the constraints introduced by the task, referent annotation achieved almost perfect inter-annotator agreement (99.7 raw agreement).

The corpus contains 1736 target expressions, of which 221



DF view of the virtual world, displayed on the DG's monitor



Part of the 2D map given to the DG showing the current room

	Session 4, 28 min 5 sec - dialog transcript	Referent Annotation
DG:	you can currently see **three buttons**... there's actually **a fourth button that's kind of hidden**	set(Button6, Button7, Button8) Button5
DF:	yeah	
DG:	by **this cabinet on the right**	Cabinet8
DF:	I know, yeah	
DG:	ok, um, so what you wanna do is you want to go in and you're gonna press **one of the buttons that's on the right hand wall**, so you wanna go all the way straight into the room and then face the wall	Button6
DF:	mhm	
DG:	there with **the two buttons**	set(Button5,Button6)
DF:	yep	
DG:	um and you wanna push **the one that's on the left**	Button6

Figure 2: Sample dialog fragment and accompanying video frame (Session 4, 28 min 5 sec). The noun phrases identifying buttons, doors and cabinets are indicated in bold.

are **AllVague** (vague + abandoned), 45 are **Empty**, and 228 are **Set**s. Table 1 presents the distribution of the various determiners and head values in a subset of the annotated expressions (only the Direction-Giver expressions not vague, ambiguous or sets).

5. Conclusions and Future Work

The release of this corpus[2] will provide a common dataset for researchers studying situated language, multimodal interaction and mobile applications. This corpus differs from previously released corpora in the wealth of information it provides: audio/video recordings of the interactions synchronized with positional/gaze information of the player and dialog transcripts aligned at word level, an xml encoding of the world information (containing bounding box information, effects associated with triggers, etc.), and referring expression annotation, associating object identifiers with noun phrases in the transcripts.

Unlike studies on negotiated reference (Brennan and Clark, 1996), the objects were not hard to describe in isolation, but because they appeared in contexts with multiple identical distractors, the partners sometimes required multiple turns to reach agreement on which item the expressions were referring to. The materials were designed this way to encourage the use of spatial relations in instructions.

This corpus may also prove useful to the artificial intelligence planning community because the tasks can be divided into clearly defined steps with preconditions and effects. Using virtual environments for language evaluation has been received well by the community and generated a recent proposal for a challenge in instruction-giving in virtual environments as an evaluation testbed for natural language generation (Byron et al., 2007).

This corpus represents the second in a line of related VR-based corpora from our lab; each version of the corpus enables different types of language technology development. However, we believe that these corpora can have utility beyond the intended usage: for example, one can also use the speech data for automatic speech recognition development. We are currently developing a third version of this corpus with a variety of native and non-native English accents in order to better understand the types of phonetic variation

[2] available online at http://slate.cse.ohio-state.edu/quake-corpora/scare/

that can occur across accents within a relatively restricted but natural vocabulary domain.

6. Acknowledgments

The authors gratefully acknowledge support for this work provided in part by NSF CAREER grant IIS-064390. The opinions expressed in this work are those of the authors and not of any funding agency.

7. References

P. Boersma and D. Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

S.E. Brennan and H.H. Clark. 1996. Conceptual Pacts and Lexical Choice in Conversation. *Learning, Memory*, 22(6):1482–1493.

Donna K. Byron and Eric Fosler-Lussier. 2006. The OSU Quake 2004 corpus of two-party situated problem-solving dialogs. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC'06)*.

D. K. Byron, A. Koller, J. Oberlander, L. Stoia, and K. Striegnitz. 2007. Generating instructions in virtual environments (GIVE): A challenge and an evaluation testbed for NLG. In *Workshop on Shared Tasks and Comparative Evaluation in NLG*, Arlington, VA, USA.

Donna K. Byron. 2005. The OSU Quake 2004 corpus of two-party situated problem-solving dialogs. Technical Report OSU-CISRC-805-TR57, The Ohio State University Computer Science and Engineering Department, September.

R. Dale, S. Geldof, and J. Prost. 2003. CORAL: Using natural language generation for navigational assistance. In M. Oudshoorn, editor, *Proceedings of the 26th Australasian Computer Science Conference*, Adelaide, Australia.

M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. 2002. MATCH: An architecture for multimodal dialogue systems. In *Association for Computational Linguistics, 2002*, pages 376–383.

J. Kelleher, F. Costello, and J. Van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167(1):62–102.

M. Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Dissertation.com.

S. Lauria, G. Bugmann, T. Kyriacou, J. Bos, and E. Klein. 2001. Training personal robots using natural language instructions. *IEEE Intelligent Systems*, 16(5):2–9.

B. Pellom and K. Hacioglu. 2001. Sonic: The University of Colorado Continuous Speech Recognizer. *Techical Report# TR-CSLR-2001*, 1.

Patrick Peruch, Loic Belingard, and Catherine Thinus-Blanc, 2000. *Spatial Cognition II, LNAI 1849*, chapter Transfer of Spatial Knowledge from Virtual to Real Environments, pages 253–264. Springer-Verlag, Berlin Heidelberg.

J. Yang, W. Yang, M. Denecke, and A. Waibel. 1999. Smart sight: a tourist assistant system. In *Proceedings of the 3rd International Symposium on Wearable Computers*, pages 73–78, San Francisco, California, 18-19 October.