Referring in Installments: A Corpus Study of Spoken Object References in an Interactive Virtual Environment

Kristina Striegnitz*, **Hendrik Buschmeier**† and **Stefan Kopp**†
*Computer Science Department, Union College, Schenectady, NY
striegnk@union.edu
†Sociable Agents Group – CITEC, Bielefeld University, Germany
{hbuschme,skopp}@techfak.uni-bielefeld.de

Abstract

Commonly, the result of referring expression generation algorithms is a single noun phrase. In interactive settings with a shared workspace, however, human dialog partners often split referring expressions into installments that adapt to changes in the context and to actions of their partners. We present a corpus of human–human interactions in the GIVE-2 setting in which instructions are spoken. A first study of object descriptions in this corpus shows that references in installments are quite common in this scenario and suggests that contextual factors partly determine their use. We discuss what new challenges this creates for NLG systems.

1 Introduction

Referring expression generation is classically considered to be the problem of producing a single noun phrase that uniquely identifies a referent (Krahmer and van Deemter, 2012). This approach is well suited for non-interactive, static contexts, but recently, there has been increased interest in generation for situated dialog (Stoia, 2007; Striegnitz et al., 2011).

Most human language use takes place in dynamic situations, and psycholinguistic research on human–human dialog has proposed that the production of referring expressions should rather be seen as a process that not only depends on the context and the choices of the speaker, but also on the reactions of the addressee. Thus the result is often not a single noun phrase but a sequence of *installments* (Clark and Wilkes-Gibbs, 1986), consisting of multiple utterances which may be interleaved with feedback from the addressee. In a setting where the dialog partners have access to a common workspace, they, furthermore, carefully monitor each other's non-linugistic actions, which often replace verbal feedback (Clark and Krych, 2004; Gergle et al., 2004). The following example from our data illustrates this. *A* is instructing *B* to press a particular button.

(1) A: *the blue button*
B: [moves and then hesitates]
A: *the one you see on your right*
B: [starts moving again]
A: *press that one*

While computational models of this behavior are still scarce, some first steps have been taken. Stoia (2007) studies instruction giving in a virtual environment and finds that references to target objects are often not made when they first become visible. Instead interaction partners are navigated to a spot from where an easier description is possible. Garoufi and Koller (2010) develop a planning-based approach of this behavior. But once their system decides to generate a referring expression, it is delivered in one unit.

Thompson (2009), on the other hand, proposes a game-theoretic model to predict how noun phrases are split up into installments. While Thompson did not specify how the necessary parameters to calculate the utility of an utterance are derived from the context and did not implement the model, it provides a good theoretical basis for an implementation.

The GIVE Challenge is a recent shared task on situated generation (Striegnitz et al., 2011). In the GIVE scenario a human user goes on a treasure hunt in a virtual environment. He or she has to press a series of buttons that unlock doors and open a safe. The challenge for the NLG systems is to generate instructions in real-time to guide the user to the goal. The instructions are presented to the user as written text, which

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 12–16,
Utica, May 2012. ©2012 Association for Computational Linguistics

means that there is less opportunity for interleaving language and actions than with spoken instructions. While some systems generate sentence fragments in certain situations (e.g., *not this one* when the user is moving towards the wrong button), instructions are generally produced as complete sentences and replaced with a new full sentence when the context changes (a strategy which would not work for spoken instructions). Nevertheless, timing issues are a cause for errors that is cited by several teams who developed systems for the GIVE challenge, and generating appropriate feedback has been an important concern for almost all teams (see the system descriptions in (Belz et al., 2011)). Unfortunately, no systematic error analysis has been done for the interactions from the GIVE challenges. Anecdotally, however, not reacting to signs of confusion in the user's actions at all or reacting too late seem to be common causes for problems. Furthermore, we have found that the strategy of replacing instructions with complete sentences to account for a change in context can lead to confusion because it seems unclear to the user whether this new instruction is a correction or an elaboration.

In this paper we report on a study of the communicative behavior of human dyads in the GIVE environment where instead of written text instruction givers use unrestricted spoken language to direct instruction followers through the world. We find that often multiple installments are used to identify a referent and that the instruction givers are highly responsive to context changes and the instruction followers' actions. Our goal is to inform the development of a generation system that generates object descriptions in installments while taking into account the actions of its interaction partner.

2 A corpus of spoken instructions in a virtual environment

Data collection method The setup of this study was similar to the one used to collect the GIVE-2 corpus of typed instructions (Gargett et al., 2010). Instruction followers (IFs) used the standard GIVE-2 client to interact with the virtual environment. Instruction givers (IGs) could observe the followers' position and actions in the world using an interactive map, and they were also provided with the same 3D view into the scene that the IFs saw on their screen. Differently from the normal GIVE-2 scenario, the IGs did not type their instructions but gave spoken instructions, which were audio recorded as well as streamed to the IFs over the network. A log of the IFs' position, orientation and actions that was updated every 200ms was recorded in a database.

Participants were recruited in pairs on Bielefeld University's campus and received a compensation of six euros each. They were randomly assigned to the roles of IG and IF and were seated and instructed separately. To become familiar with the task, they switched roles in a first, shorter training world. These interactions were later used to devise and test the annotation schemes. They then played two different worlds in their assigned roles. After the first round, they received a questionnaire assessing the quality of the interaction; after the second round, they completed the Santa Barbara sense of direction test (Hegarty et al., 2006) and answered some questions about themselves.

Annotations The recorded instructions of the IGs were transcribed and segmented into utterances (by identifying speech pauses longer than 300ms) using Praat (Boersma and Weenink, 2011). We then created videos showing the IGs' map view as well as the IFs' scene view and aligned the audio and transcriptions with them. The data was further annotated by the first two authors using ELAN (Wittenburg et al., 2006).

Most importantly for this paper, we classified utterances into the following types:

(i) **move** (MV) – instruction to turn or to move
(ii) **manipulate** (MNP) – instruction to manipulate an object (e.g., press a button)
(iii) **reference** (REF) – utterance referring to an object
(iv) **stop** – instruction to stop moving
(v) **warning** – telling the user to not do something
(vi) **acknowledgment** (ACK) – affirmative feedback
(vii) **communication management** (CM) – indicating that the IG is planning (e.g., *uhmm, just a moment, sooo* etc.)
(viii) **negative acknowledgment** – indicating a mistake on the player's part
(ix) **other** – anything else

A few utterances which contained both move and press instructions were further split, but in general we picked the label that fit best (using the above list as a precedence order to make a decision if two labels fit equally well). The inter-annotator agreement for utterance types was $\kappa = 0.89$ (Cohen's kappa), which is considered to be very good. Since the categories were of quite different sizes (cf. Table 1), which may skew the κ statistic, we also calculated the kappa per category. It was satisfactory for all ‘interesting’ categories. The agreement for category REF was $\kappa = 0.77$ and the agreement for *other* was $\kappa = 0.58$. The kappa values for all other categories were 0.84 or greater. We reviewed all cases with differing annotations and reached a consensus, which is the basis for all results presented in this paper. Furthermore, we collapsed the labels *warning*, *negative acknowledgment* and *other* which only occurred rarely.

To support a later more in depth analysis, we also annotated what types of properties are used in object descriptions, the givenness status of information in instructions, and whether an utterance is giving positive or negative feedback on a user action (even if not explicitly labeled as *(negative) acknowledgment*). Finally, information about the IF’s movements and actions in the world as well as the visible context was automatically calculated from the GIVE log files and integrated into the annotation.

Collected data We collected interactions between eight pairs. Due to failures of the network connection and some initial problems with the GIVE software, only four pairs were recorded completely, so that we currently have data from eight interactions with four different IGs. We are in the process of collecting additional data in order to achieve a corpus size that will allow for a more detailed statistical analysis. Furthermore, we are collecting data in English to be able to make comparisons with the existing corpus of written instructions in the GIVE world and to make the data more easily accessible to a wider audience. The corpus will be made freely available at `http://purl.org/net/sgive-corpus`.

Participants were between 20 and 30 years old and all of them are native German speakers. Two of the IGs are male and two female; three of the IFs are female. The mean length of the interactions is 5.24 minutes (SD = 1.86), and the IGs on average use 325 words (SD = 91).

Table 1 gives an overview of the kinds of utterances used by the IGs. While the general picture is similar for all speakers, there are statistically significant differences between the frequencies with which different IGs use the utterance types ($\chi^2 = 78.82, p \leq 0.001$). We did not find a significant differences (in terms of the utterance types used) between the two worlds that we used or between the two rounds that each pair played.

Table 1: Overall frequency of utterance types.

utterance type	count	%
MV	334	46.58
MNP	66	9.21
REF	65	9.07
stop	38	5.30
ACK	92	12.83
CM	97	13.53
other	25	3.49

Table 2: Transitional probabilities for utterance types.

	MV	MNP	REF	stop	ACK	CM	other	IF press
MV	.53	.08	.06	.06	.15	.08	.03	.00
MNP	.02	.03	.09	.02	.02	.02	.02	.80
REF	.00	.33	.19	.02	.14	.00	.02	.30
stop	.47	.03	.18	.03	.03	.16	.11	.00
ACK	.64	.08	.09	.03	.01	.10	.00	.05
CM	.53	.05	.10	.08	.01	.18	.05	.00
other	.44	.04	.12	.12	.08	.16	.00	.04
IF press	.21	.01	.00	.01	.36	.36	.04	.00

3 How instruction givers describe objects

We now examine how interaction partners establish what the next target button is. Overall, there are 76 utterance sequences in the data that identify a target button and lead to the IF pressing that button. We discuss a selection of seven representative examples.

(2) IG: *und dann drückst du den ganz rechten Knopf den blauen* (and then you press the rightmost button the blue one; MNP)
IF: [goes across the room and does it]

In (2) the IG generates a referring expression identifying the target and integrates it into an object manipulation instruction. In our data, 55% of the target buttons (42 out of 76) get identified in this way (which fits into the traditional view of referring expression generation). In all other cases a sequence of at least two, and in 14% of the cases more than two, utterances is used.

The transitional probabilities between utterance types shown in Table 2 suggest what some common patterns may be. For example, even though *move* instructions are so prevalent in our data, they are uncommon after *reference* or *manipulate* utterances.

Instead, two thirds of the *reference* utterances are followed by object manipulation instruction, another reference or an acknowledgement. In the remaining cases, IFs press a button in response to the reference.

(3) IG: *vor dir der blaue Knopf* (in front of you the blue button; REF)
IF: [moves across the room toward the button]
IG: *drauf drücken* (press it; MNP)

(4) IG: *und auf der rechten Seite sind zwei rote Knöpfe* (and on the right are two red buttons; REF)
IF: [turns and starts moving towards the buttons]
IG: *und den linken davon drückst du* (and you press the left one; MNP)

In (3) and (4) a first *reference* utterance is followed by a separate *object manipulation* utterance. While in (3) the first reference uniquely identifies the target, in (4) the first utterance simply directs the player's attention to a group of buttons. The second utterance then picks out the target.

(5) IG: *dreh dich nach links etwas* (turn left a little; MV)
IF: [turns left] there are two red buttons in front of him (and some other red buttons to his right)
IG: *so, da siehst du zwei rote Schalter* (so now you see two red buttons; REF)
IF: [moves towards buttons]
IG: *und den rechten davon drückst du* (and you press the right one; MNP)
IF: [moves closer, but more towards the left one]
IG: *rechts* (right; REF)

Stoia (2007) observed that IGs use *move* instructions to focus the IF's attention on a particular area. This is also common in our data. For instance in (5), the IF is asked to turn to directly face the group of buttons containing the target. (5) also shows how IGs monitor their partners' actions and respond to them. The IF is moving towards the wrong button causing the IG to repeat part of the previous description.

(6) IG: *den blauen Schalter* (the blue button; REF)
IF: [moves and then stops]
IG: *den du rechts siehst* (the one you see on your right; REF)
IF: [starts moving again]
IG: *den drücken* (press that one; MNP)

Similarly, in (6) the IG produces an elaboration when the IF stops moving towards the target, indicating her confusion.

(7) IG: *und jetzt rechts an der* (and now to the right on the; REF)
IF: [turns right, is facing the wall with the target button]
IG: *ja . . . genau . . . an der Wand den blauen Knopf* (yes . . . right . . . on the wall the blue button; ACK, REF)
IF: [moves towards button]
IG: *einmal drücken* (press once; MNP)

In (7) the IG inserts affirmative feedback when the IF reacts correctly to a portion of his utterance. As can be seen in Table 2, *reference* utterances are relatively often followed by affirmative feedback.

(8) IF: [enters room, stops, looks around, ends up looking at the target]
IG: *ja genau den grünen Knopf neben der Lampe drücken* (yes right, press the green button next to the lamp; MNP)

IGs can also take advantage of IF actions that are not in direct response to an utterance. This happens in (8). The IF enters a new room and looks around. When she looks towards the target, the IG seizes the opportunity and produces affirmative feedback.

4 Conclusions and future work

We have described a corpus of spoken instructions in the GIVE scenario which we are currently building and which we will make available once it is completed. This corpus differs from other corpora of task-oriented dialog (specifically, the MapTask corpus (Anderson et al., 1991), the TRAINS corpus (Heeman and Allen, 1995), the Monroe corpus (Stent, 2000)) in that the IG could observe the IF's actions in real-time. This led to interactions in which instructions are given in installments and linguistic and non-linguistic actions are interleaved.

This poses interesting new questions for NLG systems, which we have illustrated by discussing the patterns of utterance sequences that IGs and IFs use in our corpus to agree on the objects that need to be manipulated. In line with results from psycholinguistics, we found that the information necessary to establish a reference is often expressed in multiple installments and that IGs carefully monitor how their partners react to their instructions and quickly respond by giving feedback, repeating information or elaborating on previous utterance when necessary.

The NLG system thus needs to be able to decide when a complete identifying description can be given in one utterance and when a description in installments is more effective. Stoia (2007) as well as Garoufi and Koller (2010) have addressed this question, but their approaches only make a choice between generating an instruction to move or a uniquely identifying referring expression. They do not consider cases in which another type of utterance, for instance, one that refers to a group of objects or gives an initial ambiguous description, is used to draw the attention of the IF to a particular area and they do not generate referring expressions in installments.

The system, furthermore, needs to be able to interpret the IF's actions and decide when to insert an acknowledgment, elaboration or correction. It then has to decide how to formulate this feedback. The addressee, e.g., needs to be able to distinguish elaborations from corrections. If the feedback was inserted in the middle of a sentence, if finally has to decide whether this sentence should be completed and how the remainder may have to be adapted.

Once we have finished the corpus collection, we plan to use it to study and address the questions discussed above. We are planning on building on the work by Stoia (2007) on using machine learning techniques to develop a model that takes into account various contextual factors and on the work by Thompson (2009) on generating references in installments. The set-up under which the corpus was collected, furthermore, lends itself well to Wizard-of-Oz studies to test the effectiveness of different interactive strategies for describing objects.

Acknowledgments This research was supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence in 'Cognitive Interaction Technology' (CITEC) and by the Skidmore Union Network which was funded through an ADVANCE grant from the National Science Foundation.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34:351–366.

Anja Belz, Albert Gatt, Alexander Koller, and Kristina Striegnitz, editors. 2011. *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

Paul Boersma and David Weenink. 2011. Praat: doing phonetics by computer. Computer program. Retrieved May 2011, from http://www.praat.org/.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:62–81.

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

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 corpus of giving instructions in virtual environments. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2401–2406, Valletta, Malta.

Konstantina Garoufi and Alexander Koller. 2010. Automated planning for situated natural language generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1573–1582, Uppsala, Sweden.

Darren Gergle, Robert E. Kraut, and Susan R. Fussell. 2004. Action as language in a shared visual space. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, pages 487–496, Chicago, IL.

Peter A. Heeman and James Allen. 1995. The Trains 93 dialogues. Technical Report Trains 94-2, Computer Science Department, University of Rochester, Rochester, NY.

Mary Hegarty, Daniel R. Montello, Anthony E. Richardson, Toru Ishikawa, and Kristin Lovelace. 2006. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34:151–176.

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

Amanda Stent. 2000. The Monroe corpus. Technical Report 728/TN 99-2, Computer Science Department, University of Rochester, Rochester, NY.

Laura Stoia. 2007. *Noun Phrase Generation for Situated Dialogs*. Ph.D. thesis, Graduate School of The Ohio State University, Columbus, OH.

Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Mariët Theune. 2011. Report on the second second challenge on generating instructions in virtual environments (GIVE-2.5). In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 270–279, Nancy, France.

Will Thompson. 2009. *A Game-Theoretic Model of Grounding for Referential Communication Tasks*. Ph.D. thesis, Northwestern University, Evanston, IL.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1556–1559, Genoa, Italy.