The GIVE-2 Corpus of Giving Instructions in Virtual Environments

Andrew Gargett*, Konstantina Garoufi*, Alexander Koller*, Kristina Striegnitz†

* Saarland University, Saarbrücken, Germany
{gargett,garoufi,koller}@mmci.uni-saarland.de

† Union College, Schenectady, NY
striegnk@union.edu

Abstract

We present the GIVE-2 Corpus, a new corpus of human instruction giving. The corpus was collected by asking one person in each pair of subjects to guide the other person towards completing a task in a virtual 3D environment with typed instructions. This is the same setting as that of the recent GIVE Challenge, and thus the corpus can serve as a source of data and as a point of comparison for NLG systems that participate in the GIVE Challenge. The instruction-giving data we collect is multilingual (45 German and 63 English dialogues), and can easily be extended to further languages by using our software, which we have made available. We analyze the corpus to study the effects of learning by repeated participation in the task and the effects of the participants' spatial navigation abilities. Finally, we present a novel annotation scheme for situated referring expressions and compare the referring expressions in the German and English data.

1. Introduction

Understanding and generating natural-language instructions in a situated environment is a problem that has received significant attention in the past few years (MacMahon et al., 2006; Stoia et al., 2006; Zukerman et al., 2009). Most recently, the Challenge on Generating Instructions in Virtual Environments (GIVE; Byron et al. (2009)) has attracted considerable interest in the Natural Language Generation (NLG) community. GIVE is a shared task in which NLG systems must generate real-time instructions that guide a user in a virtual world. It offers novel possibilities of exploring the linguistic and non-linguistic issues involving situated NLG, while supporting a solid approach to NLG system evaluation.

In this paper, we present the GIVE-2 Corpus, a new corpus of human instruction giving in virtual environments, which is designed to support the development of NLG systems for the GIVE Challenge. We collected 45 German and 63 American English written discourses in which one subject guided another in a treasure hunting task in a 3D world. This corpus exhibits varied instruction-giving behavior, and can thus serve both as a source of data and as a point of comparison for GIVE NLG systems. We illustrate on some examples that interesting information can be extracted from the corpus, especially if phenomena of interest can be annotated by hand.

Our corpus goes beyond previous related work, such as the SCARE corpus (Stoia et al., 2008), in that it includes demographic features and spatial cognition scores, allows us to study learning effects by asking the same pair of subjects to give each other instructions repeatedly on different environments, and represents instructions and information about the virtual environment conveniently in the same data structure. It is also unique in that we use the portable, open-source software designed for GIVE to collect data, which makes collecting instructions for further languages easy in the future.

2. Method

Our task setup involved pairs of human partners, each of whom played one of two different roles. The "instruction follower" (IF) moved about in the virtual world with the goal of completing a treasure hunting task, but had no knowledge of the map of the world or the specific behavior of objects within that world (such as, which buttons to press to open doors or the safe). The other partner acted as the "instruction giver" (IG), who was given complete knowledge of the world, but had no ability to move around in it or interact with it directly, and therefore had to give instructions to the IF to guide him/her to accomplish the task.

2.1. Participants

We collected data from 15 German speaking pairs and from 21 English speaking pairs. The participants were mostly students from one German and one US university. They were recruited in pairs, and were paid a small compensation. All 30 German speaking participants were native speakers of German—17 were female and 13 male. Of the 42 English speaking participants, 35 were native English speakers, the others self-rated their English skills as near-native or very good. 16 were female, 26 male. We also recorded the participants' age, computer expertise and use of video games, and, for the English speaking participants, their college major.

2.2. Materials and infrastructure

We built the corpus collection upon the GIVE-2 software infrastructure.[1] The IF used the same 3D client program (Fig. 1) that participants in the GIVE-2 evaluation used; we replaced the automated NLG system by a graphical interface for the IG (Fig. 2). The IG interface displayed an interactive map of the world and a window for typing instructions. The map was updated in real time showing the IF's current position and orientation, and the IG could inspect

[1] http://www.give-challenge.org/research/page.php?id=software

Figure 1: The view of the virtual environment, as displayed on the IF's monitor.

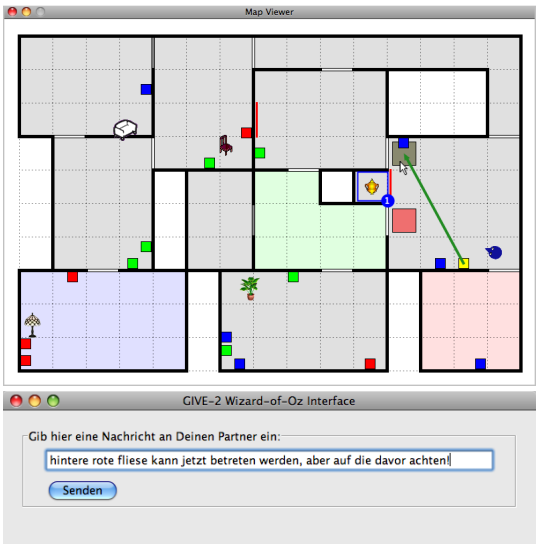

Figure 2: The IG's graphical interface, including the map interface showing the same scene as in Fig. 1 and a separate window for typing instructions.

the objects in the world by hovering the mouse over their icons. We also mirrored the contents of the IF's screen on a second screen next to the IG's monitor.

We designed three different virtual environments, shown in Figures 3–5. These were engineered so as to elicit a certain range of relevant behaviors in terms of navigation and referring expressions. As in GIVE, solving the treasure hunt required pressing a sequence of buttons in the right order to open a safe, and then picking up a trophy from the safe. In each world, the task could be solved with ten such object manipulations.

Additionally, we asked each subject to fill in the Santa Barbara Sense of Direction (SBSOD) questionnaire (Hegarty et al., 2006), a standardized self-report scale measuring the subject's ability to orient themselves in an environment. It consists of 15 items, which we translated into German for the German speaking participants.

2.3. Procedure

Each pair of subjects participated in three rounds of games, one for each world. To control for learning effects, we presented the worlds in different orders. The subjects were randomly assigned roles (IG or IF) for the first round, and switched roles after that. We gave all subjects time to get used to the software tools, and, in the IG's case, to familiarize themselves with the world before starting the round. For each round, we collected a range of data using the standard logging process of GIVE.

2.4. The GIVE-2 Corpus

The GIVE-2 corpus consists of the collected game logs, which record the experimental session with enough detail to allow for a smooth replay. Specifically, we logged the total time games took, all the instructions sent by the IG, and all the actions performed by the IF. Furthermore, we logged the IF's position and orientation every 200 milliseconds, making it possible to extract information about their movements in response to instructions and other events.

Table 1 shows a shortened and simplified excerpt of a game log. Each line starts with a timestamp. The first line indicates the IF's position and orientation and which objects he or she can see at that moment (which, in this particular example, is none). Status messages of this form are stored every 200 milliseconds. The next line shows that the IG sent an instruction. This is followed by more status messages before the IF acts, as indicated by the last line. This action message conveys that button `b10` was pressed and that the effect of this button press was to bring the safe `s1` from state `safe-state-2` into state `safe-state-3`.

Upon request interested researchers can obtain the complete logs from us together with the demographic information and the SBSOD scores that we collected from all participants. Furthermore, we provide a tool to play back experimental sessions from the game logs. This tool is also available through the Internet, such that replays of the data can be viewed online.[2]

3. Analysis

The German corpus obtained in this way consists of 2763 instructions, spread over 45 documents representing the individual rounds. On average, each round contained 61.4 instructions (SD = 24.0) and took about 752 seconds (SD = 245.8); the IF performed 12.1 object manipulations per round on average (SD = 2.6). For the English corpus, there were 63 rounds consisting of 3417 instructions. Rounds consisted on average of 54.2 (SD = 20.4) instructions, and took about 553 seconds (SD = 178.4), with the IF carrying out 11.8 (SD = 4.9) object manipulations. The numbers are on the same order of magnitude across the two languages, with the exception of completion time, which we discuss below.

3.1. Task performance factors

During both the English and German data collection experiments, subjects completed all rounds successfully. However, we can still compare the factors that influenced their performance in terms of time, number of object manipula-

[2] `http://www.give-challenge.org/research/page.php?id=give-2-corpus`

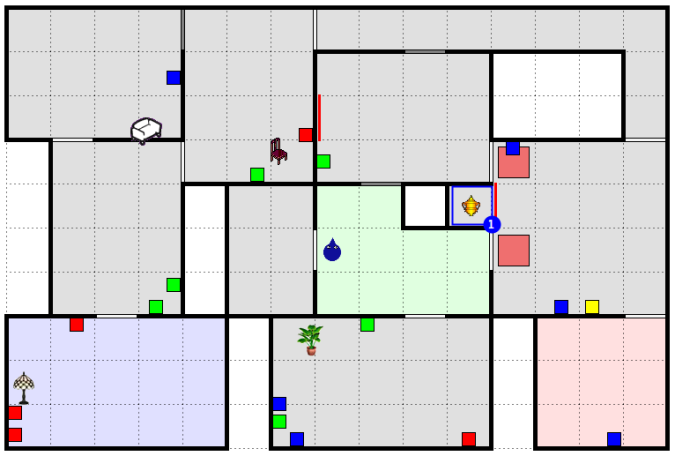

Figure 3: Virtual world 1.

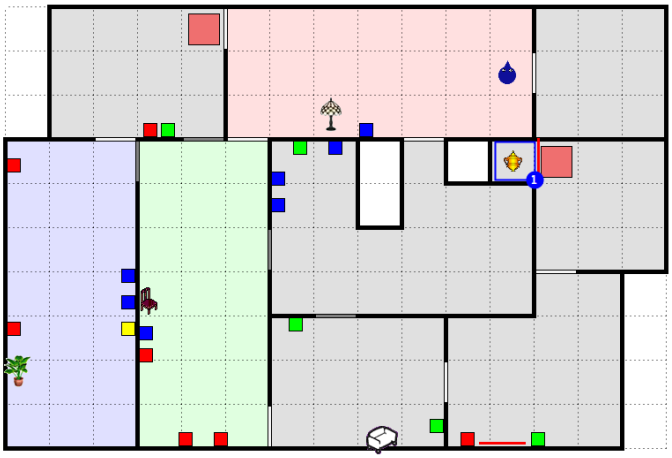

Figure 4: Virtual world 2.

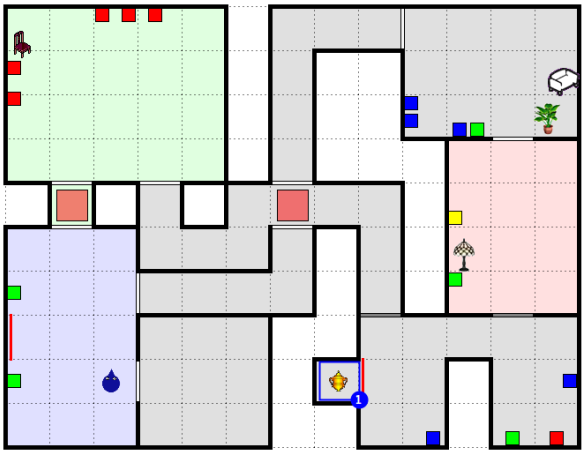

Figure 5: Virtual world 3.

```
13:18:50.020   Receive:  [Status:  pos (1.51,1.0,6.92), facing [0.03,0.0,0.99], visible:  []]
13:18:50.053   Send:     [Text:  hinten blauen knopf drücken]
13:18:50.210   Receive:  [Status:  pos (1.51,1.0,6.92), facing [0.03,0.0,0.99], visible:  []]
               ...(more status messages every 200 ms) ...
13:18:52.426   Receive:  [Status:  pos (3.45,1.0,7.66), facing [0.44,0.0,0.89], visible:  [p2, b10]]
13:18:52.509   Receive:  [Action Message:  manipulate(b10,s1,safe-state-2,safe-state-3)]
```

Table 1: A shortened and simplified excerpt of a recorded game log.

Measure	World		
	1	2	3
Time (sec)	710 (197.5)	812 (262.0)	733 (276.0)
Object manipulations	11.6 (2.1)	13.5 (3.0)	11.3 (2.2)*
Distance travelled by IF	159.8 (42.5)	193.9 (90.9)	164.7 (37.6)
Number of instructions	59.9 (27.1)	71.5 (26.6)	52.9 (13.1)*
Total instruction length	951.6 (346.6)	1030.1 (371.8)*	918.2 (410.6)
Average word length	5.3 (0.3)	5.1 (0.4)*	5.3 (0.4)

Table 2: For the German corpus, variation across worlds ($N = 15$), showing means and standard deviations. The asterisk * indicates that difference to left neighbor is significant at $p < .05$ (using Paired Samples T-test).

Measure	World		
	1	2	3
Time (sec)	560 (160.2)	557 (212.8)	543 (166)
Object manipulations	12.2 (6.0)	11.3 (3.1)	12.0 (5.4)
Distance travelled by IF	188.4 (77.4)	195.6 (124.1)	191.4 (79.0)
Number of instructions	54.7 (21.6)	54.8 (23.6)	53.2 (16.5)
Total instruction length	859.6 (230.5)	902.1 (440.7)	884.8 (386.7)
Average word length	4.8 (0.2)	4.5 (0.5)*	4.7 (0.2)

Table 3: For the English corpus, variation across worlds ($N = 21$), showing means and standard deviations. The asterisk * indicates that difference to left neighbor is significant at $p < .05$ (using Paired Samples T-test).

tions, etc., to gain some insight into what makes a given instruction discourse successful.

Unlike earlier corpus collection efforts for situated communication, we can correlate the performance measures with demographic information and spatial cognition scores. In both the German and the English data, pairs with male IGs completed the task faster ($p < .01$, ANOVA) than pairs with female IGs. In addition, we found in the German data that female IGs gave significantly longer instructions ($p < .05$) and used significantly longer words ($p < .01$), and that male IFs performed significantly more actions ($p < .05$), and covered a greater distance ($p < .05$). Note that for the English data, female IGs also gave marginally longer instructions (i.e. $p < .1$).

We also found that SBSOD scores had some effect on task performance. In the German data, short task completion times correlate with high SBSOD scores for the IG, and in the English data, IFs with high SBSOD scores travelled significantly less distance ($p < .05$), and received, and so seemed to require, significantly fewer instructions ($p < .05$). Finally, we found that our German translation of the SBSOD questionnaire was internally reliable ($N = 30$, coefficient $\alpha = .81$). With respect to demographic data and SBSOD scores for IGs and IFs, we found no other significant differences for the measures we took of the data (these measures are listed in Tables 2 and 3).

Another factor that influences task performance is which world was used in a round. Table 2 shows that German IGs gave considerably more instructions in World 2 than in the other worlds, and IFs needed more object manipulations to complete the task. These differences are much less pronounced in the English data (Table 3), just like the numbers are lower overall for the English data. One notable exception is the distance traveled, which tends to be the same or higher for US subjects than for German ones; that is, US subjects moved through the virtual worlds at a higher speed than German subjects.

We believe that these differences can be accounted for by the fact that the two cohorts had very different levels of experience with video games: Where German subjects, who were recruited from a database of previous subjects of psycholinguistic experiments and distributed quite evenly over the entire university, reported an average of 1.2 hours of video game playing per week, recruitment for the US experiment focused on engineering departments, and subjects reported an average of 6.7 hours per week of game play. We found that the time spent on video games was correlated highly significantly with lower completion times ($p < .001$), and it is reasonable to assume that subjects with video game experience will tend to explore the virtual world more on their own, and require fewer instructions to complete the task. Note also that the proportion of male participants was higher in the American cohort than in the German one (62% vs. 43%), and this had significant effects on completion times as well.

3.2. Learning effects

The fact that each pair of subjects played three rounds of games with different worlds also makes the corpus a useful resource for studying learning effects in instruction giving, beyond just the use for the GIVE Challenge. As Tables 4 and 5 show, such learning effects are indeed present. For example, the time needed to solve the task drops significantly from round B to round C in both the German and the English data. This shows that IGs learn how to give effective instructions with experience. (Remember that the IG did not change between rounds B and C.)

In the German data, the time needed to solve the task also drops from round A to round B, and additionally, IGs use significantly fewer words in round B than in round A. Since the participants switched their roles after round A, this means that the German IGs even benefited from experience

Measure	Round		
	A	B	C
Time (sec)	882 (300.5)	738 (161.2)*	636 (200.4)*
Object manipulations	12.5 (2.5)	11.5 (1.8)	12.3 (3.3)
Distance travelled by IF	177.0 (56.8)	168.0 (31.5)	173.4 (89.1)
Number of instructions	68.7 (23.7)	60.7 (25.2)	54.7 (22.4)
Total instruction length	1200.5 (441.3)	869.7 (224.3)*	829.8 (316.6)
Average word length	5.3 (0.3)	5.2 (0.3)	5.3 (0.5)

Table 4: For the German corpus, variation across rounds ($N = 15$). The asterisk * indicates that difference to left neighbor is significant at $p < .05$ (using Paired Samples T-test).

Measure	Round		
	A	B	C
Time (sec)	563 (128.0)	600 (231.0)	498 (152.3)*
Object manipulations	11.5 (4.8)	12.1 (4.2)	11.9 (5.7)
Distance travelled by IF	183.7 (57.0)	206.8 (129.5)	184.8 (85.4)
Number of instructions	50.1 (18.6)	55.5 (20.7)	57.1 (22.2)
Total instruction length	918.1 (264.2)	920.3 (431.1)	808.1 (365.3)
Average word length	4.7 (0.2)	4.8 (0.2)	4.6 (0.5)

Table 5: For the English corpus, variation across rounds ($N = 21$). The asterisk * indicates that difference to left neighbor is significant at $p < .05$ (using Paired Samples T-test).

gained while acting as the IF. In the English data, we do not see these learning effects from round A to round B. One possible explanation is that English participants, who tended to be more experienced with video games, benefited less from this "training round" about interaction with virtual worlds.

4. Annotating referring expressions

In addition to the above analysis, we can extract more detailed information from our corpus by selectively annotating phenomena of interest. To illustrate this, we annotated a sample of the corpus for the different strategies IGs used to refer to objects—a crucial ability of any NLG system. To gain some cross-linguistic insights into the task, we annotated both the German and the English sessions of World 2 (see Fig. 4).

For the annotation we devised the following scheme, which largely draws from the body of linguistic literature on spatial frames of reference (Levinson, 2003).

- **Taxonomic property:** Reference to the type of the object; e.g. "button", "square".
- **Absolute property:** Reference to a property of the object that can be determined without comparing it with other objects; e.g. "red", "yellow".
- **Relative property:** Reference to a property of the object in relation to other similar objects; e.g. "first", "middle".
- **Viewer-centered:** Reference to the object's location relative to the viewer's; e.g. "on the left", "behind you".
- **Micro-level landmark intrinsic:** Reference the object's location in relation to another object of a different kind that is movable; e.g. "by the chair", "next to the plant".
- **Distractor intrinsic:** Reference to the object's location in relation to another object of the same kind (i.e. distractor); e.g. "next to the yellow button", "closest to the blue one".
- **Macro-level landmark intrinsic:** Reference to the object's location in relation to an immovable feature of the room (i.e., the room itself, parts of the room such as corners and walls, built-in objects such as doors, safes and alarm tiles); e.g. "on the wall", "in that room".
- **History of interaction:** Reference to the object by using elements of the interaction history (i.e. previous events in the session); e.g. "from before", "as last time".
- **Visual focus:** Reference to the object by using the visual context and distinguishing visible from non-visible objects; e.g. "this one", "that".
- **Deduction by elimination:** Reference to the object by specifying which objects are not meant and letting the viewer deduce the intended one; e.g. "not that", "other one".

In total we annotated 205 REs in the 15 German and 241 REs in the 21 English sessions of our sample. All data were annotated by two annotators (a student of computational linguistics and one of the authors). The annotators worked independently, however in a second phase adjudicated any conflicts caused by errors or misunderstandings.

Strategy for RE	Frequency (%)	
	German	English
Taxonomic property (type; e.g. "button")	53.66	58.51
Absolute property (color; e.g. "red")	85.37	92.53
Relative property (e.g. "first", "middle")	6.83	4.56
Viewer-centered (e.g. "on the left", "behind you")	15.61	12.45
Micro-level landmark intrinsic (e.g. "by the chair")	13.17	17.84
Distractor intrinsic (e.g. "next to the yellow button")	10.73	14.11
Macro-level landmark intrinsic (e.g. "on the wall")	6.83	4.15
History of interaction (e.g. "from before", "as last time")	4.39	7.05
Visual focus (e.g. "this one", "that")	0.49	3.73
Deduction by elimination (e.g. "not that", "other one")	0.98	3.32

Table 6: Annotation results for RE strategies in the German and English sessions of World 2 (Fig. 4).

They thus reached agreement of $\kappa = 0.986$ in the German and $\kappa = 0.975$ in the English dataset, as indicated by Cohen's kappa coefficient. This amounts to almost perfect agreement, which is not surprising given our strictly specified annotation scheme.

One challenge in the annotation effort was identifying the REs in the situated corpus, as IGs occasionally referred to objects without specifying any of their properties; e.g. "yes", "hit". We chose not to include such instances in the annotation. Since the scheme does not consist of mutually exclusive features, we obtained the final annotation results reported here by merging the data of the two annotators.

As the results in Table 6 show, the types of referring strategies taken up by German and English speakers are comparable. A clear majority of the referring expressions in both languages involve an absolute property (color) of the target object, even at the cost of introducing redundancy. References that exploit the visual context from the viewer's perspective ("on the left"), or the location of the target with respect to distractors ("next to the yellow button") or landmarks ("by the chair") are also frequent. The fact, on the other hand, that only slightly more than half of the REs mention the basic type ("button") of the referent, seems specific to the task at hand, in which the only type of object that was ever manipulated was buttons, and is a matter for further investigation.

5. Conclusion

We have presented the GIVE-2 Corpus, a new multilingual corpus of written instructions in virtual environments. This corpus can be used to study situated communication in general, and is designed to support developing systems for the GIVE Challenge in particular. It combines, for the first time, natural-language instructions with detailed information about IF behavior and demographic data within the same resource, and opens up new possibilities of cross-linguistic research on situated language.

While we believe that even the unannotated corpus can be a useful resource, we anticipate that much more can be gained from it by selective annotation. In particular, it would be interesting to learn to generate appropriate REs and successful navigation strategies for specific contexts. We leave such efforts for future work.

6. References

D. Byron, A. Koller, K. Striegnitz, J. Cassell, R. Dale, J. Moore, and J. Oberlander. 2009. Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proc. 12th ENLG*.

M. Hegarty, D. R. Montello, A. E. Richardson, T. Ishikawa, and K. Lovelace. 2006. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34:151–176.

S. C. Levinson. 2003. *Space in Language and Cognition*. Cambridge University Press.

M. MacMahon, B. Stankiewicz, and B. Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proc. 21st AAAI*.

L. Stoia, D. M. Shockley, D. K. Byron, and E. Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proc. 4th INLG*.

L. Stoia, D. M. Shockley, D. K. Byron, and E. Fosler-Lussier. 2008. SCARE: A Situated Corpus with Annotated Referring Expressions. In *Proc. 6th LREC*.

I. Zukerman, P. Ye, K. Gupta, and E. Makalic. 2009. Towards the interpretation of utterance sequences in a dialogue system. In *Proc. 10th SIGDIAL*.