Interactive voice-based direction giving agent

DIANA PERSICO and LUCIE GALLAND*

One of the drawbacks of voice-based direction giving is that the only solution for a user that did not understand the instruction or is not sure of the way to take is to look at the map. This option can however be dangerous and/or be a loss of time for the user. Since the real world can be different from maps, where some intersections are less ambiguous, and GPS positions are not always accurate, this situation is quite frequent. To bypass this issue, we propose to develop a direction giving agents able to use grounding strategies to reduce this ambiguity. To this intent, we created a virtual world filled with ambiguous paths. Then we collected data from five couples of human participants, pairing a first person who took the instructions to a second person who led the first one along a path. Based on this dataset we created a direction giving agent which at first gives a simple and direct instruction and then can reformulate it giving further pieces of information if asked. The performances of this interactive agent were compared with the performances of a simpler agent unable to give further information. We found that our interactive agent was better perceived than a simple map based navigation application.

Additional Key Words and Phrases: direction giving agent, grounding, interaction

ACM Reference Format:
Diana Persico and Lucie Galland. 2022. Interactive voice-based direction giving agent. 1, 1 (January 2022), 7 pages.

1 INTRODUCTION

Voice-based direction giving conversational agents are widely used in common life by car drivers as well as by pedestrians or bike riders [Antikainen et al. 2006; Bartie and Mackaness 2006]. However, due to GPS tracking position errors and the ambiguity of real-life routes, simple direction giving agents can induce errors, wrong turns, and hesitations which can be dangerous and make people lose time [Boye et al. 2014]. A solution to avoid ambiguity is to look at the mini-map often provided with these agents. This solution can however be dangerous for all: pedestrians, bike riders, and car drivers who are not supposed to look at their phones [Looije et al. 2007]. On the other hand, those applications do not offer the possibility to interact with the agent and ask for clarification such as “Do I turn there ?” or “Could you repeat that?”. In this paper, we propose to add grounding strategies to a direction giving agent in order to reduce ambiguity and limit the number of wrong turns without losing time looking at a map. To this intent, we first collect and analyze a corpus of human-human interaction. Using the extracted strategies, we develop an agent able to reformulate with more detail its instruction if asked to do so. This agent is then tested on a virtual environment against a simple map-based application.

2 RELATED WORK

A rich part of the literature on direction giving agents argues on the usefulness of maps even in dangerous situations. Google navigation [goo [n. d.]], one of the most widely used navigation apps, uses spoken output and maps to guide pedestrians as if they were car drivers. To improve the use of maps, Dobesova [?] also developed a voice-controlled map. However, a navigation application displaying maps on a small screen can be strenuous and confusing [Looije et al. 2007]. From this observation, Boye et al. [Boye et al. 2014] proposed a dialog-based navigation application. Their application uses grounding strategies to ensure that GPS positions, that can be noisy, are well interpreted. Another possibility for error in such applications is incomprehension or misjudgment of the distance [Boye et al. 2014]. In [Edmonds 1993] authors present a computational model for a direction-giving model able to use grounding methods to ensure user’s comprehension. To this purpose, they start from a pre-planned direction giving dialog and adding grounding strategies along the conversation to achieve mutual belief that the plan has been understood. In [Baker et al. 2008] authors studied redundant utterances in case of a non-understanding signal during a direction giving task. To this intent, they asked a dyad of participants to reach a particular point in a building and then go back to the starting point. One participant gave directions while the other followed them. The interactions were recorded and transcripted. There were two conditions: one where the dyad could see each other and one where they could only hear each other. They found more redundant utterances in a direction giving task when the listener provided verbal or non-verbal signals of non-understanding in both conditions. Indeed since in direction giving tasks a perfect understanding of the instructions is mandatory. This gives evidence of the importance of grounding in the direction-giving agent. They also found that most instructors’ instructions were reformulated in case of understanding. A grounding direction giving agent should therefore be able to reformulate and improve its instructions in case of misunderstanding. A rich literature exists on how to improve direction giving application instructions. A large part of these works relies on landmarks. In [Michon and Denis 2001] authors have looked at the importance of landmarks in such dialogs. To this intent, they conducted two studies. The first one was composed of a participant that learned a route in the streets of Paris. This participant then needed to guide a pedestrian through this same road that was unknown to him/her. They found that landmarks were used by guides more frequently at reorientation/critical nodes, but at points where several possible directions could be followed. This result showed the importance of landmarks in direction giving dialog, even if only indirectly. To prove it more directly, they conducted a second study where participants were asked to follow simple instructions for the same route and then asked where these instructions were lacking. They found that the most common suggestion was the use of 3D landmarks at

[*] Both authors contributed equally to this research.

Authors’ address: Diana Persico; Lucie Galland.

XXXX-XXXX/2022/1-ART $15.00
https://doi.org/

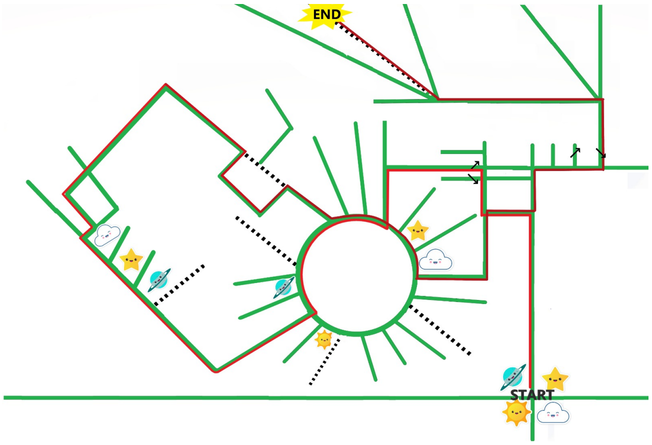


Fig. 1. Path designed with landmarks

crucial points, proving their importance in the direction-giving dialog. This study does not however address the fact that landmarks are not used at all decision points. Indeed, in some cases, relative directions are more effective as shown in [Götze and Boye 2015]. In this study, authors have asked to rate the confidence of pedestrians in relative or landmark-based instructions at different decision points. They found that in the case of simple decision points, participants were more confident in relative direction than in landmark-based direction. Based on this study we can determine which decision points should be indicated based on landmarks first hand and at which decision point landmarks should be introduced only in case of asked clarification. This study is done using a TTS and wizard of OZ setting, correcting another drawback of the first study which only looked at human-human interaction. This indicates that even with voice assistants, the use of landmarks at complicated decision points is useful. We propose in this work to find appropriate grounding strategies for comprehension of the instruction of a direction giving agent and to test whether for this particular task a dialog system is more efficient than a map-based application. Hence, in this work, we answer the two following research questions:

- **RQ1:** What grounding strategies do humans use to improve direction giving efficiency?
- **RQ2:** Can grounding strategies replace the use of a map in a direction giving agent?

3 TECHNICAL DESCRIPTION AND METHODS

3.1 Path design

We designed a path with 25 turns (see fig.1). 20 of them are simple turns, where there is no ambiguity on the way to take. The remaining 5 are conceived to be tricky (an initial crossroads where you don't know which road you are facing; 2 different turns on a roundabout with 14 exits; 2 small alleys hard to notice). Landmarks were placed along the path in strategic points, to be used in the human-human interaction to give directions. We also add coins to collect along the path and time penalty for the use of the map to account for the role play bias demonstrated by [Ewald 2012]

3.2 Human-human interaction

To answer our **RQ1**, we collected a dataset of human-human interaction related to our particular path. This corpus is collected to identify points of interest in the path and which grounding strategies are used and efficient for humans.

3.2.1 Collection setup. To create the text outputs of the assistant, we have first built a corpus of conversations of real people giving directions on our map. Six couples of participants (7 male and 5 female students of the École normale supérieure, all of them L2 English speakers) took part in the corpus data collection. We gave a copy of the map with the landmarks (fig.1) to each of the participants in the couple. We gave the map with the path only to one of them, and we asked this person to guide the other along the path. The six conversations were audio-recorded and transcribed.

3.2.2 Interpretation of the dataset. To compare the conversations, we segmented the transcripts to make correspond each part of the text to a change of direction in the path. This comparison procedure allowed us to find out about the most common grounding strategies used by the direction-giver participants to clarify previous information after a request of the direction-taker participants for confirmation, reformulation, and further information. The landmarks were used a lot for these clarifications. When able to interact with another human giving direction, participants mainly gave backchannel feedback when the instruction was understood, and they asked for reformulation or further information in case of ambiguity (see Table. 1). This allows us to answer our research question **RQ1**: The grounding strategies used by humans for efficient direction giving are mainly confirmation of understanding the instructions (about their present or next position on the map) and reformulation (more likely with addiction of with further information). From the transcripts, it emerges that the direction-takers always systematically used backchannels when the instruction was clear to them. The set of backchannels we found in the corpus is relatively small: okay, yeah, yes, uh-huh, right, sure, all right, yep, and a combination of those. In some cases, they used the repetition of the instruction as a backchannel. For our agent, we choose to use only the first strategy, using this set of backchannels, since the second one would have required language understanding. We then implemented an agent able to reformulate the direction-giving instructions with more details when asked.

3.3 Agent design

Our direction giving agent can give two versions of each instruction. A first simple one and a more complex version that helps to reduce ambiguity. The original simple instruction is given by the agent every time the participant reaches a key point, while the more complex reformulation is provided when the user asks for it.

3.3.1 Instructions design. The original instructions and reformulations of the agents were computed from our corpus of human-human interactions. In here, we looked at the direction-givers' transcripts, and we selected over the first turn inputs of the participants (namely their first attempts of giving directions) some extracts with simple directions to be used for the first turn inputs of the agent. Then, we selected, cleaned (from repairs, repetitions, and other noises), and

Example	Talker	Utterance
Confirmation of understanding through backchannels		
1	DG	And at the first intersection, you go left.
	DT	Uh-huh. Okay.
2	DG	You take the middle one left
	DT	Middle one left, right.
Request for reformulation / further information		
3	DG	So you will have to take the sixth exit.
	DT	Ok wait so.
	DG	So basically-
	DT	I cross the planet? And after the planet, the second.
	DG	Yes so you cross the planet, you don't take one of the path that touches the planet.
	DT	Uh-huh.
	DG	you take the one just after it.

Table 1. Transcript extracts

adjusted (a necessary step since our assistant is not able to understand semantic alignment or other more-refined strategies) the texts, to write the inputs of a further turn (the one coming after the participant's turn). Some examples of instructions and reformulations can be found in Table 2.

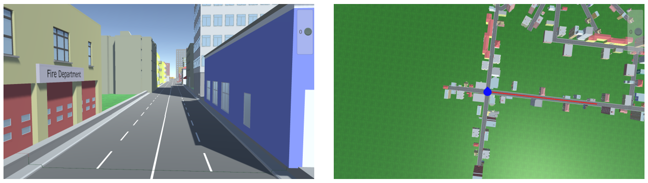
3.3.2 Implementation. The virtual environment was implemented using the Unity game engine. The generation and recognition of speech (Text To Speech and Speech To Text) are handled using Microsoft Azure software. Each time the user reaches a key point chosen in the previous section, the designed instruction is provided by the Azure TTS. A list of possible backchannels is computed using the human-human corpus collected previously. For every recognized speech, the agent checks if the input falls into the list of backchannels. The agent answers to a backchannel with another backchannel randomly picked from a list computed from the same dataset. Every recognized input that did not fall in the list of possible backchannels is considered as a request for reformulation or precision. In that case, the system provides the reformulated instruction designed in the previous section 3.3.1

4 EVALUATION

To answer our research question **RQ2**, a user study was conducted to measure the improvement in the performance and user's perception of the direction giving agent when our grounding strategy is added.

4.1 Stimuli

Participants were presented with a video game in which they were able to control a character (see fig. 2a). They were instructed on how to move in the game and how to consult the map in the trial session. They were told that they needed to collect coins in a specific order and as fast as possible. To do so, they had to follow instructions from the direction giving agent. They were also able to consult a map that displayed the path they needed to take (see fig. 2b). Accessing the map gave a penalty of 4s to the participants where they were not able to move, being displayed with a black screen. We identified one between-subject independent variable (AgentType) which has two levels. First, the baseline (base type) only gives simple instructions and can not interact with the user. The second level (interactivetype) corresponds to the setting where participants can interact with the agent. They could ask for further information when needed. A video of an example interaction is available here.



(a) View of the participant in the game

(b) View of the map by the participant

Fig. 2. Stimuli

4.2 Measurements

To measure the efficiency of the agent, we take into account the number of wrong turns and the time taken to complete the path. To measure their perception of the agent, participants were asked to rate their agreement from 1 to 5 with a list of statements (see Table.3), including participant satisfaction of the interaction (Satisfaction), whether the participant found the agent knowledgeable (Knowledgeable), whether the participant trusted the agent (Trust), whether the participant found useful to look at the map (Map), whether the participant found useful to listen to the agent's instructions (Listen), whether the participant would like to use the agent in the future (Future) and whether the participant would recommend this agent to a friend (Recommend). To control with confounding variables, participants were also asked to provide their mother tongue and their sex [Napoleon 2008] as well as answer questions about their use of assistants [Ashktorab et al. 2019] and their familiarity with videogames. They were asked to rate their agreement from 1 to 5

Turn number	Original instruction	Reformulated instruction
1	You're the starting point. You go straight.	So, you take the path between the planet and the star.
12	You go left and you take the sixth exit.	So you go left, and there is a star, and after a cloud. And so you take the one after the cloud.
19	And after... it's the third left.	So, not the one going uphill: you take the one going downhill.
21	You continue and you take the second right.	So, it's the small alley that you need to take.

Table 2. Instructions examples

Measure	Question
Satisfaction	I am satisfied with my conversation experience
Knowledgeable	I found the agent knowledgeable
Trust	I trusted the agent
Map	I found useful to look at the map
Listen	I found useful to listen to the agent's instruction
Future	I would like to use this agent in the future
Recommend	I would recommend this agent to a friend

Table 3. Questionnaire on perception of the agent

with a list of statements (see Table.4), including whether participants frequently interact with virtual agents(AssistantFreq), whether the participant enjoy talking to assistants (EnjoyAssistant), whether they find them useful (UsefulAssistant) and whether they are used to playing videogames (Videogames) .

Measure	Question
AssistantFreq	How often do you interact with one or more virtual assistants (ie Google Assistant, Siri, Alexa, ect)
EnjoyAssistant	I usually enjoy talking to assistant(s)
UsefulAssistant	The assistant(s) is/are useful
Videogames	I am used to play videogames

Table 4. Questionnaire on participant habbits

4.3 Hypotheses

We hypothesized the following:

- **H1**: The agent type (AgentType) of the conversational agent will have a main effect on the efficiency of the agent. More specifically, the participants when following the not interactive baseline agent's instructions (baselinetype) will be slower, make more wrong turns and look more at the map than the participants following the interactive agent's instructions (interactivetype).
- **H2**: The agent type (AgentType) of the conversational agent will have a main effect on the perceived quality of the interaction. More specifically, the interactions when using not interactive baseline type (baselinetype) will be perceived as worse than the interactions when the agent is interactive (interactivetype).

4.4 Results

We collected data from 11 participants from different backgrounds. 36% of them were women, 64% were men. They were all fluent in English and originated from various countries 36% were. For 36% of them their L1 was French, 45% originally spoke Italian, 1 German and 1 Farsi. Participants were assigned randomly to one of our 2 conditions. Participants were assigned randomly to one of our 2 conditions.

4.4.1 Controls. To compute the assistant score, we first compute the Cronbachs alphas score on the 3 related questions (AssistantFreq, EnjoyAssistant, UsefulAssistant) and found bad reliability ($\alpha = 0, 54$). However, using only the questions (AssistantFreq) and (EnjoyAssistant) we found acceptable reliability ($\alpha = 0, 61$)). Therefore, for each participant, we compute the assistant score by averaging the score obtained for these 2 questions. A t-test showed that participants were significantly not equally distributed among the two conditions according to the assistant score ($p = 0.05$). In condition (interactivetype), the mean score is 2,3 and in condition (baselinetype) the mean is 3,6. Therefore, we need to control for this variable in our model.

There is a difference on averadge in sex repartition over conditions but the t-test showed that this difference was not significative ($p = 0.84$). Finally, a t-test showed that the video game abilities differences across conditions is not significant ($p = 0.36$). In condition (interactivetype), participants rated them self on average at 3,8 while (baselinetype) participants' rated them self at 2,83

4.4.2 Efficiency. Our first metric for efficiency is the time taken to complete the path (see fig.3). A t-test shows that although the time taken is on average smaller for the interaction condition (535s against 570s), the difference is not significant ($p = 0, 67$). To account for our control variables, we run 2 linear regressions :

- Regression 1 :

$$\text{time} = \alpha_1 \text{condition} + \alpha_2 \text{assistant score} + \beta$$

- Regression 2:

$$\text{time} = \alpha_1 \text{condition} + \alpha_2 \text{assistant score} + \alpha_3 \text{Videogames} + \beta$$

where condition = 1 for (interactive type) and 0 for (baselinetype). The results are visible in Table 5. In both cases, the $p - value$ for condition are close to 1 ($p = 0, 96$ for regression 1 and $p = 0, 93$ for

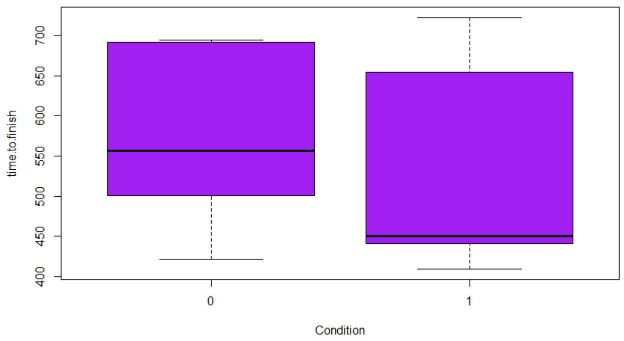


Fig. 3. Time taken to finish the path accross conditions

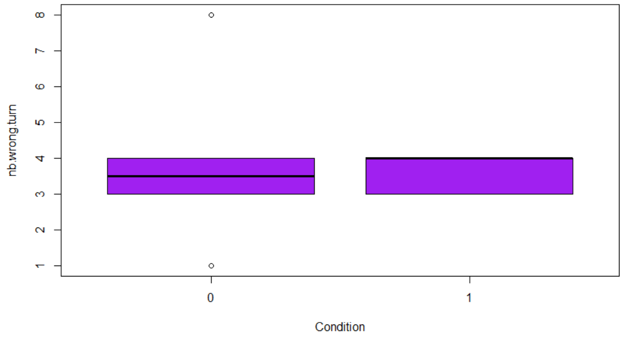


Fig. 4. Number of wrong turns across conditions

regression 2). Therefore we can not conclude on the effect of the condition on the time taken to realize the experiment.

Name	α_1	p_1	α_2	p_2	α_3	p_3
Regression 1	4,46	0,96	28,68	0,5	-	-
Regression 2	8,95	0,93	22,15	0,64	-13,86	0,65

Table 5. Linear regression on the total time

A second metric we can use to measure efficiency is the number of wrong turns. (see fig.4). A t-test shows that although the number of wrong turn is on average smaller for the interactive condition (3,6s against 3,83s), the difference is not significant ($p = 0,82$). To account for our control variables, we again run 2 linear regressions :

- Regression 1 :

$$\text{wrong turns} = \alpha_1\text{condition} + \alpha_2\text{assistant score} + \beta$$

- Regression 2:

$$\text{wrongturns} = \alpha_1\text{condition} + \alpha_2\text{assistant score} + \alpha_3\text{Videogames} + \beta$$

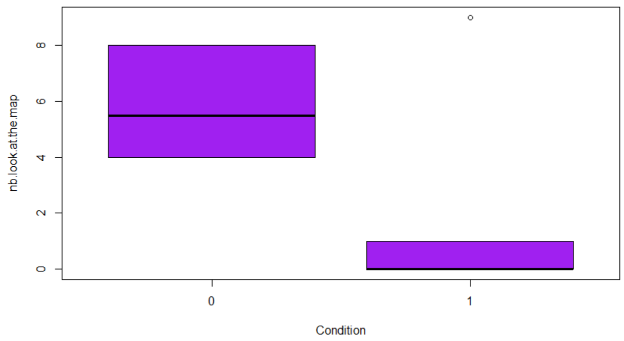


Fig. 5. Number of looks at the map across conditions

where condition = 1 for (interactive type) and 0 for (baselinetype). The results are visible in Table 6. In both cases, even if the results are not significant, the model predict a diminution of the number of wrong turn of approximately 1 with the (interactivetype) condition compare to the (baselinetype) condition.

Name	α_1	p_1	α_2	p_2	α_3	p_3
Regression 1	-1,26	0,34	-0,75	0,2	-	-
Regression 2	-1,08	0,37	-1,01	0,08	-0,55	0,13

Table 6. Linear regression on the number of wrong turns

A third and last measure for efficiency is the number of look at the map. (see fig.5). A t-test shows that the number of looks at the map is significantly smaller at the 90% level for the interactive condition (2 against 5,83) with $p = 0,1$. To account for our control variables, we also run 2 linear regressions :

- Regression 1 :

$$\text{looks at the map} = \alpha_1\text{condition} + \alpha_2\text{assistant score} + \beta$$

- Regression 2:

$$\text{looks at the map} = \alpha_1\text{condition}+\alpha_2\text{assistant score}+\alpha_3\text{Videogames}+\beta$$

where condition = 1 for (interactive type) and 0 for (baselinetype). The results are visible in Table 7. In both cases, even if the results are not significant, the model predict a diminution of the number of look at the map of approximately 2 with the (interactivetype) condition compare to the (baselinetype) condition. Due to our small number of participants, we did not find any significant effects. However the

Name	α_1	p_1	α_2	p_2	α_3	p_3
Regression 1	-2,03	0,38	1,31	0,19	-	-
Regression 2	-1,83	0,42	1,02	0,33	-0,61	0,35

Table 7. Linear regression on the number of looks at the map

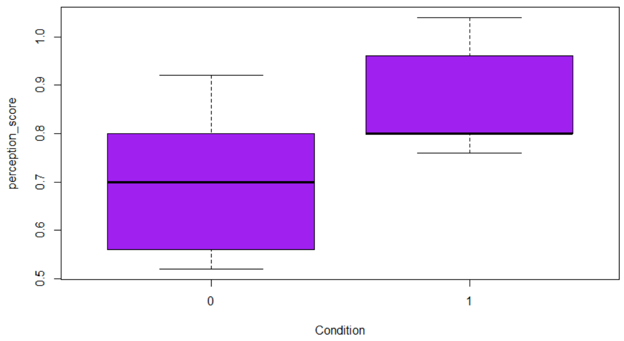


Fig. 6. Perception score across conditions

results are coherants with our hypothesis **H1** : the time taken is on average smaller in the (interactive type) condition compare to the (baselinetype) condition and the number of wrong turns and looks at the map tend to be smaller in the (interactivetype) condition even when we control with the confounding variables.

4.4.3 Perception score. To compute the perception score, we first compute the Cronbachs alphas score on the 7 questions about perception (Satisfaction, Knowledgeable,Map,Listen, Trust,Future and Recommend)and found good reliability (α =0,7). For each participant, we compute the perception score by averaging the score obtained for each question with the Map question counting as a negative score. A t-test showed a significant difference at the 90% level across conditions for the perception score (see fig.**??**) with a mean score of 0,87 for the (interactivetype) condition and of 0,7 for the (baselinetype) condition. To account for our control variables, we also run 2 linear regressions :

- Regression 1 :

$$\text{perception score} = \alpha_1\text{condition} + \alpha_2\text{assistant score} + \beta$$

- Regression 2:

$$\text{perception score} = \alpha_1\text{condition}+\alpha_2\text{assistant score}+\alpha_3\text{Videogames}+\beta$$

where condition = 1 for (interactive type) and 0 for (baselinetype). The results are visible in Table 8. In both cases, the linear regressions show a significant impact at the 95 % of the condition on the perception score. The (interactivetype) condition is perceived approximately 30% better than the (baseline type) condition. Therefore our hypothesis **H2** is verified

Name	α_1	p_1	α_2	p_2	α_3	p_3
Regression 1	0,27	0,02	0,07	0,01	-	-
Regression 2	0,26	0,02	0,09	0,04	0,04	0,15

Table 8. Linear regression on the perception of the agent

5 CONCLUSION

In this article, we collected a small corpus of human-human interactions to determine the commonly used grounding strategies. From this corpus we were able to answer our **RQ1**: reformulation is commonly used by humans to correct a misunderstanding of the guide. A direction-giving agent for a virtual environment was then created based on the before mentioned corpus. This agent can interact with the user and reformulate its direction instruction when asked to. We finally conducted a user study that answered our **RQ2**: Users perceive as better an interactive direction giving agent compared to a map-based navigation application. The results also tend towards showing that an interactive direction giving agent might be more efficient than a map-based application but a more extensive study is necessary to significantly prove it.

5.1 Limitations

The game was not designed to handle large deviations from the path, which led to issues during the experiment when the participant did not follow the instructions. Some participants also could not finish the experiment because of motion sickness. Another limitation was that our agent is only able to reformulate once each instruction. The input of the participant is also classified in only two categories, the response of the agent could be more specific with a language understanding module. Another limitation was that many participants did not feel comfortable talking to the agent and preferred to look at the map even though this was associated with a time penalty

5.2 Future Works

In Future Work, we could cover the limitations by adding a language understanding module and a more specific response from the agent. The game can also be improved to avoid technical issues. Another way to improve our agent would be to add to it the second grounding strategy emerging from the human-human interaction corpus: confirmation. A way to proceed could be to analyze the number of backchannels given by the user and react to a diminution of their frequency, which could signal misunderstanding or hesitation.

REFERENCES

[n. d.]. Google Inc.: Google maps navigation. . http://www.google.com/mobile/navigation/.

Harri Antikainen, Jarmo Rusanen, Sami Vartiainen, Mauri Myllyaho, Jari Karvonen, Markku Oivo, Jouni Similä, and Kari Laine. 2006. Location-based Services as Tool for Developing Tourism in Marginal Regions. *Nordia Geographical Publications* 35 (12 2006), 39–50.

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. *Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

Rachel E. Baker, Alastair J. Gill, and Justine Cassell. 2008. Reactive Redundancy and Listener Comprehension in Direction-Giving. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (Columbus, Ohio) *(SIGdial '08).* Association for Computational Linguistics, USA, 37–45.

Phil J. Bartie and William A. Mackaness. 2006. Development of a Speech-Based Augmented Reality System to Support Exploration of Cityscape. *Trans. GIS* 10 (2006), 63–86.

Johan Boye, Morgan Fredriksson, Jana Götze, Joakim Gustafson, and Jürgen Königsmann. 2014. *Walk This Way: Spatial Grounding for City Exploration.* 59–67. https://doi.org/10.1007/978-1-4614-8280-2_6

Philip Edmonds. 1993. A Computational Model of Collaboration on Reference in Direction-Giving Dialogues.

Jennifer D. Ewald. 2012. "can you tell me how to get there?": Naturally-occurring versus role-play data in direction-giving. *Pragmatics* 22, 1 (2012), 79–102. https://doi.org/10.1075/prag.22.1.03ewa

Jana Götze and Johan Boye. 2015. "Turn Left" Versus "Walk Towards the Café": When Relative Directions Work Better Than Landmarks. https://doi.org/10.1007/978-3-319-16787-9_15

Rosemarijn Looije, Guido Brake, and Mark Neerincx. 2007. Usability engineering for mobile maps. 532–539. https://doi.org/10.1145/1378063.1378150

Pierre-Emmanuel Michon and Michel Denis. 2001. When and Why Are Visual Landmarks Used in Giving Directions?. In *Spatial Information Theory*, Daniel R. Montello (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 292–305.

Siena Napoleon. 2008. From Here to There: A Sociolinguistic Study in Gender and Direction-Giving. *Indiana Undergraduate Journal of Cognitive Science* 2 (08 2008).