*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 600–609
Marseille, 11–16 May 2020


# RDG-Map: A Multimodal Corpus of Pedagogical Human-Agent Spoken Interactions

**Maike Paetzel**[1], **Deepthi Karkada**[2], **Ramesh Manuvinakurike**[3]
[1]Uppsala University Social Robotics Lab, Sweden
[2]Intel Corp, [3]Intel Labs, Hillsboro, Oregon, United States
maike.paetzel@it.uu.se, deepthi.karkada@intel.com, ramesh.manuvinakurike@intel.com

**Abstract**

This paper presents a multimodal corpus of 209 spoken game dialogues between a human and a remote-controlled artificial agent. The interactions involve people collaborating with the agent to identify countries on the world map as quickly as possible, which allows studying rapid and spontaneous dialogue with complex anaphoras, disfluent utterances and incorrect descriptions. The corpus consists of two parts: 8 hours of game interactions have been collected with a virtual unembodied agent online and 26.8 hours have been recorded with a physically embodied robot in a research lab. In addition to spoken audio recordings available for both parts, camera recordings and skeleton-, facial expression- and eye-gaze tracking data have been collected for the lab-based part of the corpus. In this paper, we introduce the pedagogical reference resolution game (RDG-Map) and the characteristics of the corpus collected. We also present an annotation scheme we developed in order to study the dialogue strategies utilized by the players. Based on a subset of 330 minutes of interactions annotated so far, we discuss initial insights into these strategies as well as the potential of the corpus for future research.

**Keywords:** Multimodal Corpus, Conversational Games, Serious Dialogue Games, Rapid Spoken Dialogue, Human-Robot Interaction, Crowd-sourcing

## 1. Introduction

In this paper, we present the RDG-Map corpus, an extension of the previously developed Rapid Dialogue Game (RDG) series (Paetzel et al., 2014; Zarrieß et al., 2016). The core dynamic of the RDG domain is to score as high as possible in a spoken dialogue game under time pressure, which leads to spontaneous rapid responses from the players. Due to its importance in daily conversations, the goal of rapid dialogue games is to capture related phenomena of spontaneous natural speech in numerous tasks of varying complexity. Such corpora can then be used for building state-of-the-art dialogue processing models that can make dialogue systems more efficient and human-like (Paetzel et al., 2015; Manuvinakurike et al., 2017).

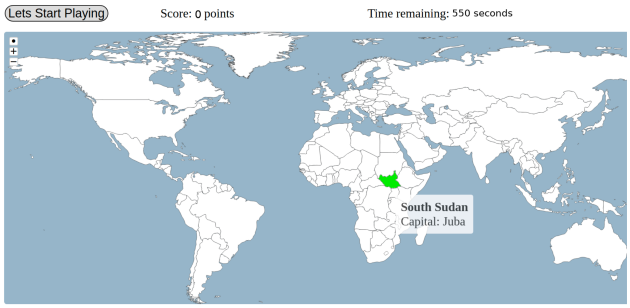
In comparison to the other games developed as part of the RDG series, RDG-Image, RDG-Phrase, and RDG-Pento, finding countries on the world map requires a more complex description strategy than it is required for identifying images, phrases and Pentomino pieces. In these simpler domains, players followed straightforward dialogue strategies only involving the current target. In RDG-Map, on the contrary, we observe more complex dialogue interactions making use of other countries and past context. Such strategies result in rich conversations that can be an ideal platform for research on a variety of dialogue-related topics, such as language understanding, generation, and dialogue management. Figure 1 shows a sample dialogue in this domain.

The RDG-Map game contributes a new task that *captures complex interactions involving references to information grounded in past turns*. Introducing pedagogical aspects into the RDG-domain also allows us to *expand the relevance of our domain to real-world applications*. RDG-Map was specifically designed to increase people's geographic literacy while still being fun over the course of repeated interactions. Geographic literacy is an important skill in an ever-increasing globalized world that the population at large is lacking (CFK and Geographic, 2006; CFR and Geographic, 2016), and we could already show in previous work that the RDG-Map game increases people's self-assessed geographic skills (Paetzel and Manuvinakurike, 2019).

This paper not only presents the RDG-Map domain and the related game dynamics in detail (Section 3); we *primarily discuss characteristics and preliminary findings from a large data collection involving 34.8 hours of spoken dialogue conversations* (Section 4). For this data collection, we developed a Wizard of Oz interface to remote-control an agent playing the game. This technique allowed us to collect an authentic corpus of people believing to play with an artificial agent without the necessity of having the full agent capabilities developed. The corpus we collected consists of two parts: The first initial 48 game sessions were recorded on the web with an unembodied agent. With the experience gained from this data collection, we revised the game design and control interface and expanded the corpus in a large lab-based data collection involving 58 individuals playing 161 game sessions with an embodied agent. Due to the adaptions made between the two parts of the corpus collection, the parts are not meant to be comparable but to complement each other.

The second contribution of this paper is the *introduction of a dialogue annotation scheme* that we developed to capture the specific game-related description strategies as well as common dialogue features in our corpus (Section 5). We show that annotations using this scheme can be performed with a high inter-annotator agreement and we conclude the paper by presenting insights into people's approaches to describing countries on the world map (Section 6).

Information about the corpus including the annotation scheme is available online at `https://rdglearn.com/corpus`. The corpus will be made available for download at this link as well.




**Dialogue Excerpt**
**Dir:** Look at the um Africa... top three biggest countries uh on top of Africa... You see those?
**Mat:** Yes
**Dir:** The one to the furthest right... is Egypt
**Mat:** Ok
**Dir:** Go two down... that is South Sudan
**Mat:** What does it look like?
**Dir:** Kinda looks like seahorse laying on its back
**Mat:** Got it

Figure 1: The map as presented to the Director. The target country is highlighted in green and information about the country is presented when hovering over it. Below the map is a sample conversation between the human Director (Dir) and the agent Matcher (Mat).

## 2. Related Work

Dialogue in the context of vision has been of interest for a long time (Clark and Wilkes-Gibbs, 1986; Tanenhaus et al., 1995). Lately, neural network based approaches for visual dialogue have shown promise in various sub-tasks which evoked interest in the community to build corpora for visual dialogue of various complexities. These visual dialogue tasks have, for instance, been modeled as reference resolution games and have been used to collect dialogue interactions between human players (e.g. (Stoia et al., 2008; Kazemzadeh et al., 2014; Paetzel et al., 2014; Zarrieß et al., 2016; De Vries et al., 2017)) as well as in human-robot teams (Skantze, 2017). Such tasks involve conversations between the interlocutors who identify one or multiple targets among distractors. Similarly, corpora for visual question answering (Antol et al., 2015; Das et al., 2017) have been developed. Such tasks involve answering the questions asked by the user about visual content through conversations. Other visual dialogue tasks include map navigation (understanding the navigation instructions in a visual scene) (Anderson et al., 1991; de Vries et al., 2018), image retrieval (Guo et al., 2018) and image manipulations (Manuvinakurike et al., 2018; Kim et al., 2019). The corpus presented in this paper includes visual reference resolution and question answering, among other tasks. Since our corpus consists of spoken interactions, we observe phenomena that are typically present in spontaneous spoken conversations and not specific to visual dialogue.

One of the aspects that sets our corpus apart from related task-oriented game-based reference resolution corpora is the pedagogic value. While Intelligent Tutoring Systems (ITS) have achieved appreciable learning gains in recent times (Lesgold et al., 1992; Freedman, 1999; Koedinger et al., 1997; Mitrovic and Ohlsson, 1999; Gertner and VanLehn, 2000; Graesser et al., 2001; Litman and Silliman, 2004; Graesser et al., 2004; McNamara et al., 2004; Craig et al., 2013; Koedinger et al., 2013; Pane et al., 2014; Graesser, 2016; Trinh et al., 2017), they usually rely heavily on experts hand-authoring the pedagogical content. In recent times learning dialogue policies automatically have been shown promising results and could potentially solve the necessity for extensive hand-authoring in ITS (Georgila et al., 2019). Learning such tutoring policies, however, requires large amounts of data. In recent years, crowdsourcing has been employed to overcome this drawback (Mitros and Sun, 2014; Baker, 2016). In this work, we present a scaleable crowd-sourcing approach to learn system responses that facilitate learning in a pedagogical reference resolution game. Paetzel and Manuvinakurike (2019) have discussed the learning aspect of the RDG-Map domain and analyzed people's self-reported increase in geographic literacy after playing the game in further detail.

## 3. Rapid Dialogue Game Map

**Game design:** RDG-Map is a spoken collaborative game in which one of the players is assigned the role of the *Matcher* and the other the role of the *Director*. The Matcher's goal is to locate countries on the world map based on the descriptions given by the Director. One of the countries is randomly selected as the *target country* and highlighted on the Director's screen. The Director is free to choose any verbal description to help the Matcher identify the target country, including saying the name of the country. The map on the Director's screen is labeled with the names of all countries in order to provide guidance when giving descriptions (cf. Figure 1). The map of the Matcher is not labeled, making the name of the country likely not sufficient to identify the target country, unless the Matcher has prior knowledge about its location. By showing the Director the name of all countries, (s)he is implicitly provided with learning content (name of the country, neighbors, continent, etc.) that can be taught to the Matcher. Teaching the Matcher the names of countries leads to future rewards since it later accelerates finding countries. The team scores a point for each correct guess made by the Matcher. The goal for the players is to identify as many countries and score as high as possible in the given 10 minutes game time. The RDG-Map corpus we present in this paper is based on a human-agent team playing the game together. For collecting the corpus under conditions as realistic as possible without fully developing an autonomous agent, the agent was remote-controlled by a human operator. The human player was always assigned the role of the Director.

**The corpus:** Collecting the corpus was divided into two parts. First, we developed a version of the game for collecting interactions over the crowd-sourcing platform Amazon Mechanical Turk (AMT)[1]. In a similar work, Manuvinakurike et al. argued that the spoken dialogue data collected using crowd-sourcing environments provide diverse and high-quality interaction at low-cost

[1]https://mturk.com

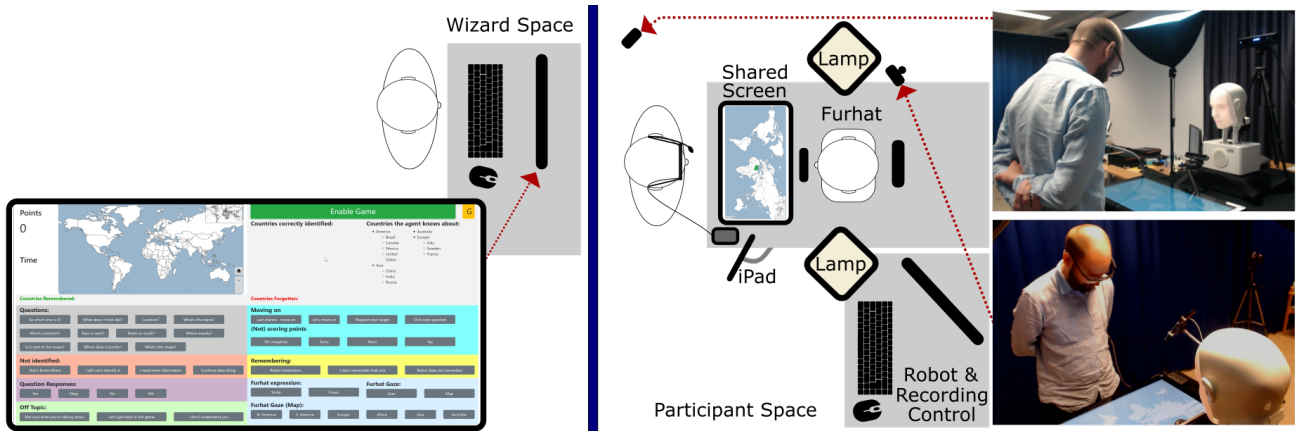



Figure 2: Setup for the lab-based data collection including the interface controlling the robot's speech, gaze and country selection (bottom left).

(Manuvinakurike and DeVault, 2015; Manuvinakurike et al., 2015). However, this comes at the disadvantage of fewer control over environmental conditions and lower quality of the audio recordings. Based on this first part of the corpus collected over the web, we implemented a revised version of the game playable with an embodied agent in a controlled laboratory environment. In the following, we describe the details of the web-based (Section 3.1) and embodied (in-lab) part of the corpus (Section 3.2).

### 3.1. Web-based Data Collection

**System:** The first version of the game was deployed online using HTML5 SimpleWebRTC[2]. 50 participants were recruited on AMT to play the game. Before being directed to the game queue, participants had to provide informed consent and demographic information and read a description of or watched a video about the game rules. All participants reported being native English speakers. To ensure participants understood the game, they needed to pass a short test containing three questions about the game rules. Since the same human operator controlled the agent for all games, participants often had to wait for several minutes before they were paired with the remote-controlled agent. Once paired, the Director could click on the 'start the game' button to begin the 10-minute long gameplay. The agent in this data-collection was represented by the female CereProc voice Kate and had no embodiment. Its actions were controlled using a custom-designed button-interface that allowed the operator to select from pre-recorded voice samples and make the country selection on behalf of the agent. The country selection of the Matcher was not visible to the Director. The agent thus needed to explicitly communicate when they selected a country so that the Director could move on to the next target. Once the Director requested the next target, both the Director and Matcher were shown whether they scored a point for the previous target country, but they were not given a chance to try again if the selection of the Matcher was wrong. The order of the target countries was the same for all participants. The operator was instructed to have prior knowledge of a small subset of countries (Australia, Brazil, Canada, China, India, Italy, Mexico, Russia, USA) based on the list of countries that more than 50% of Americans can find on the world map (CFK and Geographic, 2006).

After the game, participants were asked to fill out a questionnaire about their experience with the agent and the game. They then received monetary compensation for their participation in the study, which was independent of their final game score.

**Features:** Collecting data online has several implications on the quality of the collected audio data. While participants were strongly encouraged to wear a headset, not all of them followed this rule. Thus, the audio quality varies between speakers. For some users, significant background noises from a television or other people are present. Echos of the agent speech are also a common artifact. Similarly, we could not control the loudspeakers of the participants; hence, the intelligibility of the agent and the clarity of the agent's speech might have varied significantly between users. These factors result in a noisy but scalable, low-cost method for data collection.

### 3.2. Embodied Data Collection

**System:** For the lab-based data collection, the agent Matcher was embodied in a Furhat robot (Al Moubayed et al., 2012). Furhat is a blended embodiment consisting of a firm mask of a male face onto which a facial texture is projected from within. Since the visual appearance of the robot is very masculine, the agent's voice was switched to the male CereProc voice, William. 60 students from an international Masters course were recruited at Uppsala University to participate in the data collection. The minority of participants were native English speakers. Before starting with the game, they gave informed consent to participate in the study, read the game rules, and provided demographic information.

**Adjustments:** Based on the experience with the crowdsourced data collection certain adjustments were made to the game rules. From the data collected online, we realized that an important cause of frustration during the game was the lack of feedback on the agent's country selection. The Director in the web-version of the game could not determine the reason for non-optimal performance by the agent

[2]https://www.simplewebrtc.com/

or why they did not score a point for a given target. Since the agent was controlled by an operator that eliminated language understanding errors, the most common cause for lack of point scoring was an incorrect description given by the Director, like confusing continent names (e.g., using South America for Africa) or directions (e.g., saying east instead of west). Thus, to make the game more pleasant and increase the learning experience due to fast feedback on wrong descriptions, we decided to make the Matcher's selection visible to the Director in the updated version of the game. In this embodied version, the screen of the Matcher was shared between the two, so the Director could see the Matcher's selection.

During some pilot interactions using the updated game rules, we saw that the most common description strategy would now become very incremental: The Director gave a very broad description, the Matcher selected a country randomly in that region and then they would gradually move closer to the target (e.g. "move left"). To prevent this strategy, we limited the Matcher's selection to two per target country. As an incentive to getting the first selection correct, the team scored two points if the initial selection of the Matcher was correct and only one point if the second selection was correct. With these updated game rules, we were able to increase the transparency and collaboration within the game but to still keep the descriptive strategies comparable to our first version of the game.

Two major changes were made in the control interface for the agent between the web-based and the lab-based implementation. In the beginning of the interaction, the robot commented on being assembled in Stockholm, Sweden, and having traveled to France in previous years. Consequently, those two countries were added to the initial knowledge base of the agent. The second change makes use of the physical embodiment of the agent and allows the operator to direct the robot's gaze to the location whenever the Director mentions the name of a continent (America, Europe/Africa, or Asia/Oceania). We also added other questions and responses to resolve common game situations the agent could not appropriately react to in the web-based data collection.

**Setup:** A shared screen showing the Matcher's selection was placed on a table between the participant and the robot, as visualized in Figure 2. The Director's private screen was visible on an iPad attached to the side of the table so that it was not visible from the robot's perspective.

In the lab-based data collection, participants were recorded with several different devices: They wore a close-range Sennheiser microphone for voice recording as well as Tobii glasses for gaze tracking. Above the robot, a Microsoft Kinect was attached to record the participant's posture and allow the robot to seek eye contact with the human partner. Below the robot, a RealSense camera recorded their facial expressions. The entire scene was captured from two different angles using two Logitech webcams. All data except for the eye-tracking data are accurately synchronized. Using a beep tone played at the beginning of the data collection ensures that the eye-tracking data can be synchronized with the other recordings after the interaction.

| | $RDG-Map$ | |
|---|---|---|
| | Web | Lab |
| **Full Corpus** | | |
| Audio length (in min) | 480 | 1610 |
| # sessions | 48 | 161 |
| # unique speakers | 48 | 58 |
| Points scored | M = 14.44 | 25.9 |
| | SD = 6.44 | SD = 10.6 |
| | Min = 3 | Min = 3 |
| | Max = 29 | Max = 53 |
| Points scored/min | 1.44 | 2.59 |
| **Annotated Subset** | | |
| Audio length (in min) | 160 | 170 |
| # unique speakers | 16 | 17 |
| # word tokens | 8648 | 7919 |
| Speaking rate (words/min) | 54.05 | 46.58 |

Table 1: An overview of the Human-Agent web and lab-based corpus.

**Features** Since this data collection was part of a larger experiment, the agent initiated a short social chat before and after the gameplay. For the lab-based corpus collection, we were specifically aiming for people returning to play with the agent after multiple days of break in between sessions. This allows us to study the recurrence of individual dialogue features like the commitment to certain descriptive strategies and references. To study the influence of the agent's level of human-likeness on people's strategies when talking to the agent, the Furhat robot was equipped with three different facial textures, either showing the face of a real human, a mechanical face or a morph between the two. These textures were varied between participants, so participants always played with the same robot appearance for recurring sessions.

## 4. Corpus Characteristics

The two parts of the RDG-Map corpus we collect online and offline are summarized in Table 1.

Our RDG-Map Web part of the corpus consists of 48 unique speakers, each playing one ten minutes long session with the remote-controlled agent. Two of the original 50 speakers had to be excluded due to audio problems or quitting the experiment prior to completion. Participants scored on average 6.44 points during the game or 1.44 points per minute. However, the success in the game varied significantly: The most successful player scored 29 points (2.9 points per minute), the least successful one only 3 points (0.3 points per minute).

For the RDG-Map Lab part, we recruited 60 participants, out of which two were excluded because they suspected the agent to be remote-controlled. As described above, participants had between one and three sessions with the agent, resulting in 163 sessions in total. Two of these individual sessions were excluded from the corpus due to technical problems. In the resulting 161 sessions participants scored an average of 2.59 points per minute. The fact that participants scored almost double the number of points can be explained with the different scoring system implemented for the lab-based game that rewarded players with two points in case

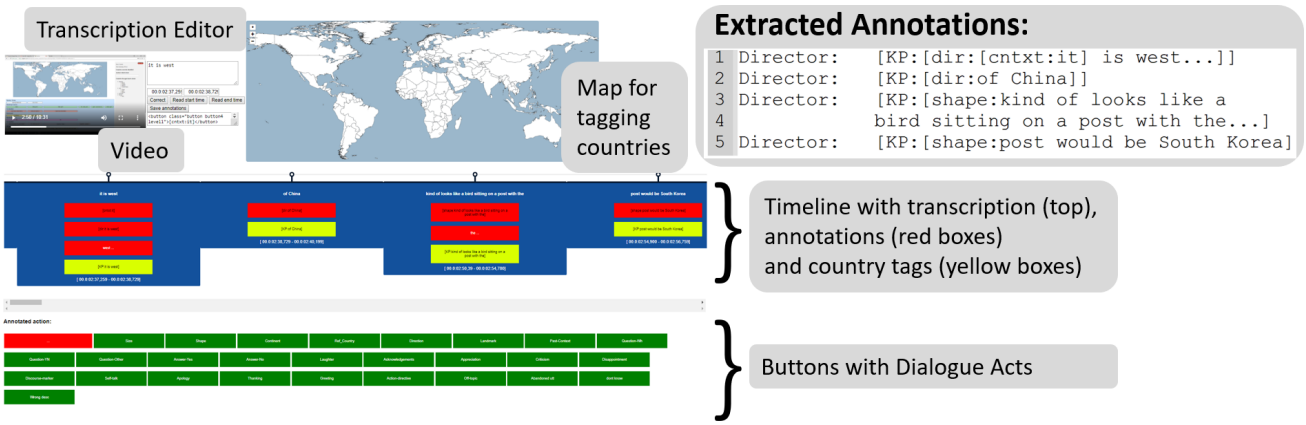



Figure 3: Custom developed tool for correcting transcriptions, annotating dialogue acts as well as labelling speech segments with the respective country.

the first guess was correct. Similarly to the web-corpus, we could observe a significant difference between the player's scores ranging from 0.3 to 5.3 points per minute.
For both the web and the lab-based part of the corpus, demographic information like self-disclosed gender, age and nationality are available for each speaker.

## 5. Transcriptions and Annotations

The two parts of the corpus were automatically transcribed using off the shelf Automatic Speech Recognition (ASR) systems and were manually corrected afterwards. Utterances by the agent and the human were tagged and processed separately. Only the human utterances were manually annotated since the agent's utterances were logged and the intentions could thus be extracted automatically. The ASR generated transcriptions were divided into inter-pausal units separated by silence and annotated individually with dialogue acts (DAs) following the annotation scheme provided in Table 2. While most of the DAs were designed specifically for our domain, some general tags motivated by Bunt (2009) were included as well.
The dialogue acts can be mainly divided into 4 categories:

1. **Target descriptions**: The human director is describing the target country so the agent can select the respective country. The descriptions in this category include the size, shape, direction, continent, landmark, anchor country description, anaphoric reference to countries already grounded in the conversation and a label for incorrect descriptions (descriptions for which the director confuses the continents or directions which lead to incorrect country identifications).

2. **Question-Answers**: The interlocutors are engaged in question-answer exchanges. The questions (Wh, Yes-no and others) by the Director are annotated as well as their yes-no answers to the questions asked by the agent. If the player did not know the answers to a question, the answer is marked as unsure (A-U).

3. **Reactions**: The human director is reacting to the agent's actions or utterances. A common reaction to an agent's utterance is a generic acknowledgement. In addition, we annotate appreciation of the agent's efforts and negative reactions to it (including criticisms and disappointments). These are important to model the user's frustration and engagement with the game and the agent.

4. **Others**: These dialogue acts include the action-directives (AD), abandoned utterances (Aband), Discourse markers (Disc-M), self-talks (Self-T), greetings (Greet) and off-topic speech (Off-Top). While we follow the definitions from Bunt (2009) to mark the labels in this category, off-topic is specific to this domain and includes all non-game utterances and descriptions which do not help the agent to identify the target country since it exceeds its knowledge (e.g: "This is the country Cristiano Ronaldo is from").

We developed a web-based annotation tool to annotate the corpus. The tool is shown in Figure 3. The text is annotated by first marking the words and then selecting the relevant DA for the utterance.
Each utterance is also annotated with the country being described by the human director. This step was necessary in spite of logging the target country as one of the common strategies to describe a target country involved locating a different, much easier 'anchor' country in close proxemics to the target county. As soon as this anchor was established, the Director would then describe the target county in relation to the anchor.
So far, we annotated a subset of 330 minutes and Table 1 shows the statistics from the annotated subset.

**Efficacy of the annotation scheme** In order to validate the annotation scheme, we performed an inter-annotator agreement (IAA) analysis on four-game interactions (approximately 12% of the annotated corpus, two from the web and two from the lab-based part). The annotations were performed independently by three annotators. The raw IAA on these four interactions was 0.857 and Fliess-kappa (N=3) was 0.676 indicating strong agreement on the annotations.

**Differences in web & lab-based data** We observe some interesting differences between the distribution of dialogue acts between the web and the lab-based part of the corpus.

| DA | % Lab | % Web | % Tot | Description | Example |
|---|---|---|---|---|---|
| **Target Descriptions** | | | | | |
| Size | 1.90 | 2.89 | 2.41 | Description of the target's size | 'it's kinda small', 'the biggest one in Africa' |
| Shape | 7.93 | 10.49 | 9.24 | Description of the target's shape | 'looks like pac-man', 'many islands' |
| Anaphora | 18.57 | 20.24 | 19.43 | Countries referenced by the director | 'The one that you had before' |
| Direction | 16.44 | 19.38 | 17.94 | Description of the target in relation to another country | 'west of Egypt', 'right below France' |
| Continent | 8.88 | 9.41 | 9.15 | Description of the target in relation to a continent | 'in Africa', 'below Europe', 'southern Africa' |
| Landmark | 4.44 | 6.04 | 5.26 | Description of the target in relation to a region or landmark on the map that is not a country or continent | 'in the Mediterranean', 'close to the ocean', 'in the Middle East' |
| Anchor | 0.32 | 3.67 | 2.03 | Additional DA for descriptions in which the director teaches the agent another country or landmark first in order to establish a starting point for the target description | Participant: ' Do you know where Egypt is?' Agent: 'No' Participant: 'Egypt is the most north-east country in Africa' |
| Wrong | 0.72 | 0.73 | 0.73 | Additional DA to mark descriptions that are incorrect | 'Venezuela is in South Africa' |
| **Questions & Answers** | | | | | |
| Q-WH | 0.27 | 0.69 | 0.49 | Question starting with what, when, where, who, whom, which, whose, why and how | 'What else do you need to know?' |
| Q-YN | 3.17 | 6.56 | 4.91 | Yes/No Question | 'Did you get it?', 'Do you know where Chad is?' |
| Q-Other | 0.27 | 0.26 | 0.27 | Any other question type | |
| A-Y | 6.57 | 0.26 | 3.34 | Yes or otherwise positive response to a question | Agent: 'Can you say more about the location?' Participant: 'Sure' |
| A-N | 2.85 | 0.13 | 1.46 | No or otherwise negative response to a question | Agent: 'Is it next to the ocean?' Participant: 'No, it's landlocked' |
| A-U | 1.09 | 0.22 | 0.64 | The director is either unsure or otherwise undecided | Agent: 'What does it look like?' Participant: 'I don't know, it's hard to say' |
| **Reactions to the agent** | | | | | |
| Ack | 3.71 | 5.78 | 4.77 | Acknowledgement of the agent's response | Agent: 'I need more information' Participant: 'Okay' |
| Apprec | 3.94 | 0.95 | 2.41 | Appreciation or thanking an agent for its action or response | 'Ok, good!', 'Very nice', 'Thank you' |
| Neg-R | 0.27 | 0.65 | 0.46 | Negative reaction to an agent's action or response such as disappointment or criticism | 'You're really dumb' |
| Apology | 0.32 | 0.35 | 0.33 | The director apologizes for the action or past utterance. | 'Ok, sorry, never mind' |
| **Other** | | | | | |
| AD | 0.63 | 1.04 | 0.84 | The director provides directives for the agent to take some action | 'Ok, let's move on', 'Click on the country' |
| Aband | 4.57 | 3.28 | 3.91 | Description that is abandoned mid-way | 'and it', 'it looks like' |
| Disc-M | 5.34 | 4.49 | 4.91 | Discourse marker | 'okay', 'well', 'let's see' |
| Self-T | 1.68 | 0.78 | 1.22 | Directors talk to themselves | 'let me see here', 'is it Italy? I think so' |
| Off-Top | 6.07 | 1.51 | 3.73 | Any description given that is either not relevant to the game or cannot be identified by the agent | 'do you know where space is?' 'it has many mountains' |
| Greet | 0.05 | 0.17 | 0.11 | These descriptions refer to greeting utterances by the director | 'Hi', 'Hello' |

Table 2: Dialogue scheme for the corpus annotation. The columns refer to the Dialogue acts labeled, percentage of DAs covered in the corpus in total and in the lab (L) and web-based part (W) separately, description of the DA and examples from the corpus.

The first set of differences can be attributed to the different conditions and populations in the lab-based corpus collection. The most notable difference is the spike in off-topic conversations that occurred in the lab environment. The majority of these can be attributed to the participants asking questions or making comments to the researcher regarding game rules or technical difficulties during the game. Since participants on Amazon Mechanical Turk had no possibility to contact a researcher directly, off-topic conversations were less common. If online-participants encountered problems, they either dropped out, or the session was automatically terminated in the web-based framework. A difference in the number of abandoned utterances can also be observed between the two parts. A likely explanation is that most participants in the lab-based data collection were non-native English speakers which decreased their fluency.

The second set of differences can be attributed to the changes made in the game rules and the enhancements of agent's utterances for the embodied agent. These changes mostly reflect in the increase in Director responses (both yes, no and undecided) as well as a decrease in yes/no questions. The most common question asked by the Director in the web-based part of the corpus is whether the agent has already selected a country. Since the Director can see the country selection in the lab-based version of the game, such questions were not present. To improve the agent's gameplay and to add more variability to its speech output, we added a number of questions that could be asked by the agent to identify the target. Consequently, the number of responses by the Director increased.

As discussed in Section 3, the main objective for adding a shared screen was for the Director to see the agent's selection and consequently decrease the number of negative reactions towards the agent. Indeed, a decrease in negative reactions and an increase in appreciation markers towards the agent could be observed in the lab-based corpus. While this could be attributed to the shared screen, it could also be due to the fact that the agent was equipped with an embodiment and a researcher being present in the room.

## 6. Descriptive Strategies

Our corpus presents interesting insights into descriptive strategies employed by the Director to help the Matcher identify the target country across different sessions. To ground the conversation around the coordinates of the target country, the Directors use directional descriptions related to the continent, other countries, or landmarks like oceans or regions. To further identify the target country, various appearance-related descriptions such as shape and size are used by the Director. In the following, we will discuss the most common descriptive strategies and point out differences between the web and lab-based parts of the corpus.

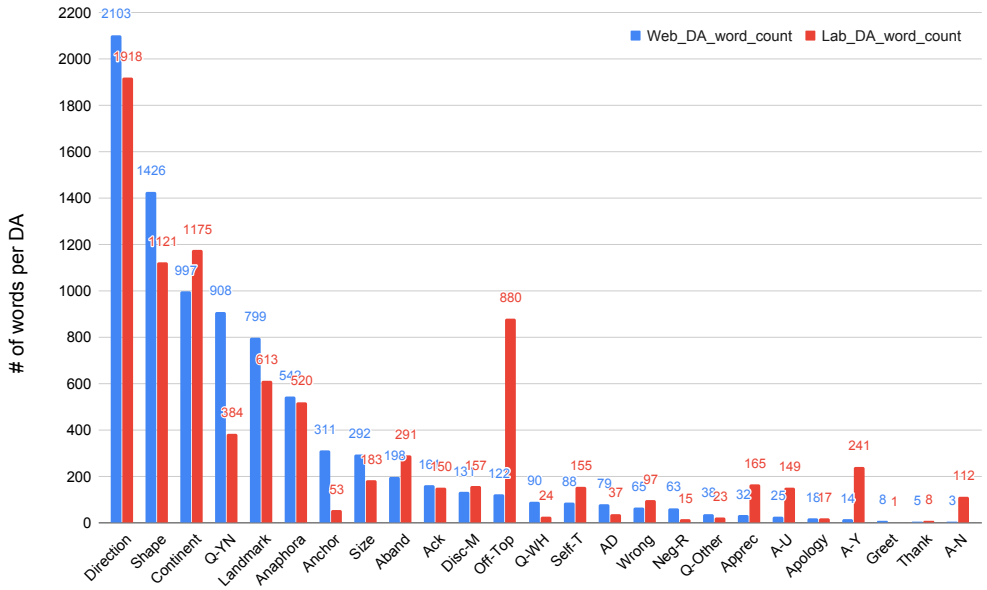



Figure 4: Word count of dialogue acts from the web and lab-based part of the corpus.

**Continent** The name of the continent or directional descriptions within the continent are often used as the first frame of reference. Directors playing with the robot in the lab generally made more use of this strategy than people playing online (cf. Fig. 4). We believe the robot directing the gaze towards continents when they were used as part of a description might have reinforced this behavior in the lab-based corpus because it served as a clear indication of the agent correctly understanding the description.

**Landmarks** Descriptions given in relation to landmarks on the world map are often as broad as continent-related descriptions. Common examples of landmarks in our corpus include world regions like the Middle East and specific bodies of water like the Mediterranean Sea, among others.

**Directions, Anchors and Anaphora** The most common descriptive strategy to aid the Matcher in identifying the target country is the usage of directive descriptions in relation to other countries, for example:

**Director** It's west of Egypt

These directional descriptions could cross several countries (e.g. describing Pakistan as "south or Russia") and are not always given using clear directional markers like "east" or "below":

**Director** If Chad just blew its nose you would see Nigeria

While country-based descriptions are the most commonly used strategy for describing a country, they are not necessarily successful given the agent's limited initial knowledge. Thus, participants sometimes describe *anchors*, other countries or landmarks on the world map that are not the target country and that are deemed to be easier to identify:

**Director** If you look at the Mediterranean Sea and you look at Africa it's the second country from the left. The first country from the left is Egypt. Second country is Libya.

Once an anchor is established, the Director can describe the target country in relation to this country. Using this logic, participants can build long reference chains, describing multiple anchoring countries before getting to the target country. Anaphora, most often referencing the original target country, are crucial in such descriptions. Co-reference resolution is thus important for the language understanding of the agent, as can be seen in the following example:

**Director** Do you know where France is?
**Matcher** Yes
**Director** Well below France, there is a country. This is not the one, it's to the left of that country.

Sometimes a Director references back to countries that had been described as targets or anchors prior to the current target description, which shows the need for saving the past context in order to resolve the current description. This makes resolving co-references in our domain an interesting and challenging problem in comparison to other domains.

**Shape** With the approximate region of the target country being established by directional descriptions, the agent Matcher was often left uncertain with a small group of potential target countries. In this case, shape descriptions were faster to identify the target and thus the agent would specifically inquire about the shape. Shape descriptions can generally be grouped into two categories: *Abstract shapes* and *associations*. Abstract shapes contain descriptions like "rectangle" for Turkey or "pointy" for Somalia. On the contrary, associations use other objects that look similar to the shape of the country as a reference. For example, Germany was frequently described as Pac-Man and Pakistan as a dinosaur. Many of the associations were recurring and thus the agent could identify the country based on it. Encouraged by their success, some Directors became very inventive when describing countries, e.g., describing Ethiopia as "Darth Vader's helmet".

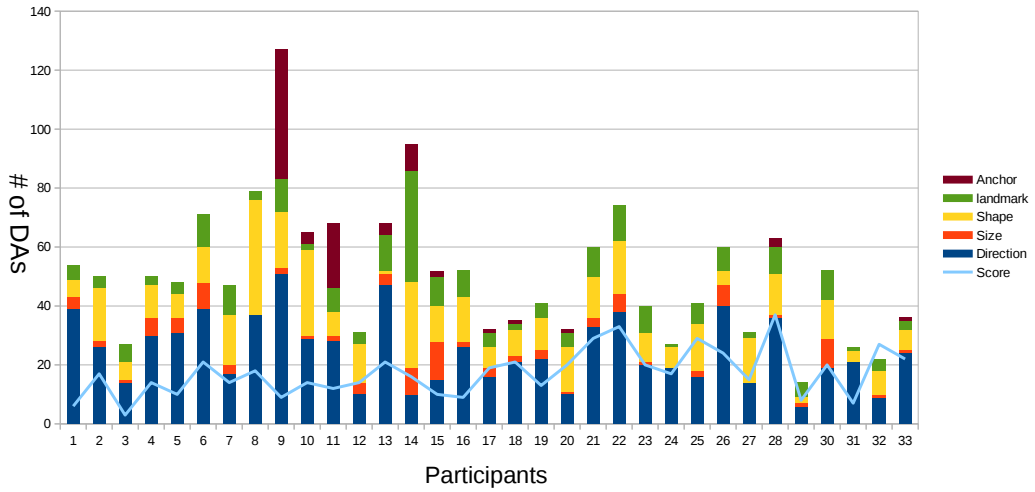



Figure 5: Distribution of descriptive strategies used by the Directors and points scored in the 33 annotated game sessions.

**Size** The size of the country is a feature that is comparably seldom used to describe the target country. One reason may be that it is not directly inquired by the agent, which may lead to many Directors not identifying it as a descriptive strategy. In order to understand how accurate the size descriptions given by the Directors are, we used K-Means to find size-based clusters for the countries in our corpus. We identified five size clusters, ranging from $< 4k\ km^2$ to $> 2M\ km^2$. The spoken descriptions given by the Director were manually grouped into five similar categories: (1) tiny or very small, (2) small, (3) medium, (4) big, (5) very big or large. The Pearson's r between the K-Means clusters on the actual size and the manual clusters on the verbal descriptions was $r = 0.49, p < .001$. This shows that people's size description generally matches the actual size of the country on the world map. However, the location of the country seems to play an important role in the accuracy of people's descriptions. Poland, for example, was consistently grouped into the largest category based on the Director's descriptions, while in reality it belongs to the smallest one. This is especially interesting since it is smaller than the neighboring country Germany, a fact often misjudged by Directors describing Poland as "the biggest" in the area.

**Comparing strategies among players** Even though the guidelines for the operator were developed such that different descriptions could lead to identifying a country, we were interested in understanding whether there was one strategy predominant across Directors who scored particularly well in the game. Figure 5 shows the distribution of descriptive strategies and the points scored by each of the 33 participants that were annotated. We found no single description strategy that significantly correlated with the points scored (Pearson coefficient $r_{shape}$=0.29, $r_{size}$=0.15, $r_{landmark}$=0.1, $r_{direction}$=0.19, $r_{anchors}$=0.21, $p > .1$ for all comparisons). This gives confidence that the success in the game did not depend on finding the winning strategy but rather on giving generally good descriptions.

## 7. Conclusion & Future work

In this work, we presented the RDG-Map corpus, a collection of dialogues between humans and a remote-controlled agent playing a pedagogical reference resolution game. We present two parts of the corpus, one that was collected online using crowd-sourcing and the other one offline in a research lab. In addition to corpus characteristics, we presented an annotation scheme specifically designed to extract descriptive strategies in our domain. Part of the corpus has already been annotated with a high inter-annotator agreement, and we discussed insights gained from these annotated dialogues. Paetzel and Manuvinakurike (2019) have already shown promising results regarding an increase in self-reported geographic literacy after playing the game with the remote-controlled agent. By implementing a fully autonomous version of the game, we hope to make this game available to the general public for them to play and increase their geography skills. This annotated corpus will help facilitate the development of automated agents that can play the game in the role of the Director and Matcher.
The corpus presented in this paper also poses a unique challenge for the language understanding of automated agents that we consider interesting for the broader research community. Such agents need to incorporate complex language understanding, dialogue management and language generation capabilities. The corpus can be used to study dialogue phenomena like rapid turn-taking, filled pauses, discourse markers and co-reference resolution over the course of multiple turns, among others.

## Acknowledgements

Part of this work was supported by the COIN project (RIT15-0133) funded by the Swedish Foundation for Strategic Research. Thanks to G. Perugia and R. Kessler for assisting with the lab-based data collection and technical setup, as well as K. Georgila for her input on the design of the study and the game.

## Bibliographical References


Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. (2012). Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7403 LNCS:114–130.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Baker, R. S. (2016). Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education*, 26(2):600–614.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

CFK, R. and Geographic, N. (2006). Geographic Literacy Survey.

CFR and Geographic, N. (2016). What College-Aged Students Know About the World, A Survey on Global Literacy.

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

Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., and Okwumabua, T. (2013). The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers & Education*, 68:495–504.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. (2017). Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., and Courville, A. (2017). GuessWhat?! Visual Object Discovery Through Multi-Modal Dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

de Vries, H., Shuster, K., Batra, D., Parikh, D., Weston, J., and Kiela, D. (2018). Talk the Walk: Navigating New York City through Grounded Dialogue. *arXiv preprint arXiv:1807.03367*.

Freedman, R. (1999). Atlas: A Plan Manager for Mixed-Initiative, Multimodal Dialogue. In *AAAI-99 Workshop on Mixed-Initiative Intelligence*, pages 1–8.

Georgila, K., Core, M. G., Nye, B. D., Karumbaiah, S., Auerbach, D., and Ram, M. (2019). Using Reinforcement Learning to Optimize the Policies of an Intelligent Tutoring System for Interpersonal Skills Training. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 737–745. International Foundation for Autonomous Agents and Multiagent Systems.

Gertner, A. S. and VanLehn, K. (2000). Andes: A Coached Problem Solving Environment for Physics. In *International Conference on Intelligent Tutoring Systems*, pages 133–142. Springer.

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., and Harter, D. (2001). Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine*, 22(4):39–39.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., and Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36(2):180–192.

Graesser, A. C. (2016). Conversations with AutoTutor Help Students Learn. *International Journal of Artificial Intelligence in Education*, 26(1):124–132.

Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., and Feris, R. (2018). Dialog-based Interactive Image Retrieval. In *Advances in Neural Information Processing Systems*, pages 678–688.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. (2014). ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798.

Kim, J.-H., Kitaev, N., Chen, X., Rohrbach, M., Zhang, B.-T., Tian, Y., Batra, D., and Parikh, D. (2019). CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., and Mark, M. A. (1997). Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education*, 8:30–43.

Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., and Nixon, T. (2013). Using Data-Driven Discovery of Better Student Models to Improve Student Learning. In *International Conference on Artificial Intelligence in Education*, pages 421–430. Springer.

Lesgold, A., Lajoie, S., Bunzo, M., and Eggan, G. (1992). SHERLOCK: A Coached Practice Environment for an Electronics Trouble-shooting Job. *Computer-Assisted Instruction and Intelligent Tutoring Systems: Shared Goals and Complementary Approaches*, pages 201–238.

Litman, D. J. and Silliman, S. (2004). ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. In *Demonstration papers at HLT-NAACL 2004*, pages 5–8. Association for Computational Linguistics.

Manuvinakurike, R. and DeVault, D. (2015). Pair Me Up: A Web Framework for Crowd-Sourced Spoken Dialogue Collection. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 189–201. Springer.

Manuvinakurike, R., Paetzel, M., and DeVault, D. (2015). Reducing the Cost of Dialogue System Training and Evaluation with Online, Crowd-Sourced Dialogue Data Collection. *Proceedings of the 19th Workshop on the*

*Semantics and Pragmatics of Dialogue (goDIAL)*, page 113.

Manuvinakurike, R., DeVault, D., and Georgila, K. (2017). Using Reinforcement Learning to Model Incrementalityin a Fast-Paced Dialogue Game. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SigDial)*, pages 331–341, Saarbrücken, Germany.

Manuvinakurike, R., Bui, T., Chang, W., and Georgila, K. (2018). Conversational Image Editing: Incremental Intent Identification in a New Dialogue Task. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SigDial)*, pages 284–295.

McNamara, D. S., Levinstein, I. B., and Boonthum, C. (2004). iSTART: Interactive Strategy Training for Active Reading and Thinking. *Behavior Research Methods, Instruments, & Computers*, 36(2):222–233.

Mitros, P. and Sun, F. (2014). Creating Educational Resources at Scale. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 16–18. IEEE.

Mitrovic, A. and Ohlsson, S. (1999). Evaluation of a onstraint-based Tutor for a Database Language. *International Journal of Artificial Intelligence in Education*, 10:238–256.

Paetzel, M. and Manuvinakurike, R. (2019). "Can you say more about the location?" The Development of a Pedagogical Reference Resolution Agent". *Dialog for Good - Workshop on Speech and Language Technology Serving Society (DiGo)*.

Paetzel, M., Racca, D. N., and DeVault, D. (2014). A Multimodal Corpus of Rapid Dialogue Games. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4189–4195, Reykjavik, Iceland.

Paetzel, M., Manuvinakurike, R., and DeVault, D. (2015). "So, which one is it?" The effect of alternative incremental architectures in a high-performance game-playing agent. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 77–86, Prague, Czech Republic.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., and Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144.

Skantze, G. (2017). Predicting and Regulating Participation Equality in Human-Robot Conversations: Effects of Age and Gender. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 196–204. IEEE.

Stoia, L., Shockley, D. M., Byron, D. K., and Fosler-Lussier, E. (2008). SCARE: a Situated Corpus with Annotated Referring Expressions. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Citeseer.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Trinh, H., Asadi, R., Edge, D., and Bickmore, T. (2017). RoboCOP: A Robotic Coach for Oral Presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):27.

Zarrieß, S., Hough, J., Kennington, C., Manuvinakurike, R., DeVault, D., Fernández, R., and Schlangen, D. (2016). PentoRef: A Corpus of Spoken References in Task-oriented Dialogues. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 125–131, Portorož, Slovenia.