Learning Task-Oriented Dialogues through Various Degrees of Interactivity

Sebastiano Gigliobianco
University of Potsdam
Potsdam, Germany
sebastiano.gigliobianco@uni-potsdam.de

Dimosthenis Kontogiorgos
Massachusetts Institute of Technology
Cambridge, MA, USA
dimos@csail.mit.edu

David Schlangen
University of Potsdam
Potsdam, Germany
david.schlangen@uni-potsdam.de

Abstract

Due to the scarcity of dialogue datasets compared to the vast amount of non-interactive text utilized in large language models, this work aimed to collect dialogues featuring referring expressions in collaborative tasks. In an interactive study, two participants were paired up and presented with the same image of a puzzle. One participant, the instruction giver, had access to an annotated version of the puzzle board, and their task was to find a description that enabled the other participant, the instruction receiver, to identify and select the referent target. The paper investigates whether and to what extent manipulations of the complexity of the task and the degree of interactivity between the users affect the type of referring language that is collaboratively constructed. The results revealed that the aforementioned manipulations had a statistically measurable impact on the type of referring expressions generated by the participants and that interactivity had a major effect on how instructions were collaboratively and iteratively refined.

1 Introduction

The ability to accurately resolve referential entities in text remains challenging. Whether in the domain of information retrieval, question answering, or machine translation, the interpretation and handling of references are fundamental to the coherence of automated text processing systems. Towards addressing these critical challenges, this work presents PentoNav: a dataset containing annotated logs of task-oriented cooperative dialogues.

We developed a collaborative task where two human users were matched in a chat room and were shown a picture of a *Pentomino* puzzle. The key difference between the two users was that one of the participants, the instruction giver (IG), had access to a labelled version of the puzzle with a bounding box around the target piece and had to describe it so that the instruction receiver (IR), who could only see the unlabelled image, could uniquely identify and select the correct piece.

The main question we examined was **whether the complexity of the task and the degree of interactivity between the users have a measurable effect on the type of referring expressions generated**. To study how participants adapt to different settings, we modified the underlying experiment along two main dimensions: *task complexity* and *interactivity between users*.

Our findings indicate that the degrees of interactivity in online interactions have a significant effect on how referring utterances are co-constructed, especially how the feedback of the listener affects the incremental production of referring expressions. Differences were also found in the instruction receiver's task accuracy and response time, as well as in the length of the referring expression produced. Overall, more complex tasks required a higher cognitive load from both participants, indicating that higher task complexity also increases the collaborative effort. Furthermore, a higher degree of interactivity degree also appeared to align with increased accuracy and longer referring expressions.

The resulting dataset (PentoNav) is a publicly available corpus containing 640 *Pentomino* puzzles and descriptions equally distributed among three complexity levels and four experiment designs. PentoNav provides valuable insights into the various strategies participants employ during the collaborative task.

2 Related Work

2.1 Referring Expression Generation

Reference is the linguistic phenomenon in which a noun phrase refers to an entity within a sentence (Stede, 2012). Recent research in the area of referring expression generation has examined how to collect referring expressions generated by humans trying to solve a common task (i.e., the *ReferItGame* (Kazemzadeh et al., 2014)). During such tasks, humans are typically shown pictures of real-world scenes, and generate referring expressions for highlighted objects (Perkins, 2021).

Other datasets combine methods from computer vision and NLP (Loáiciga et al., 2021), investigating phenomena of reference and coreference resolution in task-oriented dialogues with visual support. A lot of referring expression generation work focuses on puzzles such as the *PentoRef* (Zarrieß et al., 2016) and *Pento-DIA Ref* (Sadler and Schlangen, 2023) datasets. Both works use the *Pentomino* puzzle paradigm, that this work also utilizes. In comparison to *PentoRef*, *Pento-DIA Ref* is a synthetic dataset where expressions are generated by the incremental algorithm (Krahmer and van Deemter, 2012).

Some interesting work has been carried out in the area of instruction oriented dialogue. Notable examples contain the Tactical Speaker Identification Speech Corpus (TSID) collected by Graff et al. (1999) or the HCRC Map Task Corpus by the University of Edinburgh (1993). Both corpora feature dialogues between participants tasked with finding a route between two points on a map. Another similar experiment was conducted by Brennan et al. (2013) and differs from the previous studies in that one participant received directions by telephone while searching for target locations on the Stony Brook University campus. Once the target location was reached, the participant had to take a photograph, which was later compared with the target image described by the other participant, alongside the GPS data from the mobile phone.

2.2 Task Complexity

Many studies from the fields of linguistics and cognitive sciences have been conducted to measure the time it takes a person to resolve referential expressions. Elsner et al. (2017), for example, demonstrated how visual complexity measurably affects referring expression generation. During this study, participants were shown abstract scenes containing multiple objects that share some features and were instructed to describe the target piece. Referring expressions were extracted and analyzed, showing how visual complexity can delay or facilitate description generation.

Similarly, Clarke et al. (2013) showed how complicated and cluttered scenes translate to longer referring expressions. The study was conducted by showing participants images from the *Where's Wally* book with a bounding box surrounding the target piece, and they were tasked to write a referring expression for it. The authors were able to find a correlation between the median length of the expression and task complexity showing once again that complexity plays a crucial role when describing objects.

Another setting in which task complexity is commonly used is referential gaze modelling as shown by Alacam et al. (2022) who trained different models on the Eye4Ref work (Alacam et al., 2020) to predict whether a gaze from a participant is directed at a referent object or not. Increasing task complexity was correlated with a decreasing F1-Score.

These studies suggest that task complexity has a measurable effect on people's effort to describe common objects and construct referential expressions, which we use as one of the main dimensions to examine in this study.

2.3 Degrees of Interactivity

While the broad concept that a higher degree of interactivity between participants leads to a higher success rate has been observed in general tasks (Handzic and Low, 2002), de Weck et al. (2019) observed this concept in the field of referring expression generation. In their study, they analyzed the referring expressions of twenty parents telling a story either to their child or to an adult and found an overall wider range of referring expressions when participants talked to children. While not evaluating the strategies itself, the study showed that the interaction setting influences the type of generated referring expressions.

Dialogue is by nature incremental, which means that it's processed step by step as information is delivered (Schlangen and Skantze, 2009). This problem has already been addressed in the past, for example, by Manuvinakurike et al. (2017) who leverage reinforcement learning to incremental dialogue policy learning in dialogue games and show how this new approach outperforms a human-like baseline system in a collaborative task.

Apart from the incremental nature of human dialogue, the degree of interactivity is also relevant to what type of medium people use to produce referring expressions and how information is distributed to different channels (including the non-verbal channels). Receiving feedback in referring expression generation through backchannels or non-verbal cues has also been shown to affect how references are collaboratively produced (Kontogiorgos, 2022).

Variations in the degree of interactivity play a crucial role: changing the degree of interaction should influence the strategies adopted by people to refer to objects as they may have different ways to receive feedback. This work aims to investigate how these variations affect the production of referring expressions including how task complexity correlates with interactivity.

3 Experimental Setup

Similarly to the *Pento-DIA Ref* dataset (Sadler and Schlangen, 2023), the data collection was conducted by pairing two participants in a chat room with an image of a *Pentomino* puzzle (Figure 1). One of the participants, the instruction giver (IG) was able to see a labelled image with a highlighted piece and had to describe it to the other participant, the instruction receiver (IG), who needed to select it based on the description and the unlabelled image of the same board. To examine strategy differences across diverse settings, four variations of the puzzle's basic design were created, and the complexity of the *Pentomino* boards was modified. During the data collection, the instruction giver was not aware whether the instruction receiver was a human or a computer program.

The participants were recruited using the Prolific platform (Prolific), and the only requirements were proficiency in the English language and being at least 18 years of age. Each participant was only allowed to participate in the study once. The participants were aged between 18 and 58 (on average 27.5 with a standard deviation of 7) and mostly based in Europe. About a third of the participants declared English to be their first language. Out of the 48 participants in total, 27 reported female and 21 male. On average, each participant took 15 minutes to complete the task with a standard deviation of 8 minutes.

The data was collected using SLURK (Götze et al., 2022): an extensible chat server optimized for conducting multi-modal dialogue experiments and data collections, with a framework for creating abstract representations and interfaces to object manipulation tasks.

3.1 Task Complexity

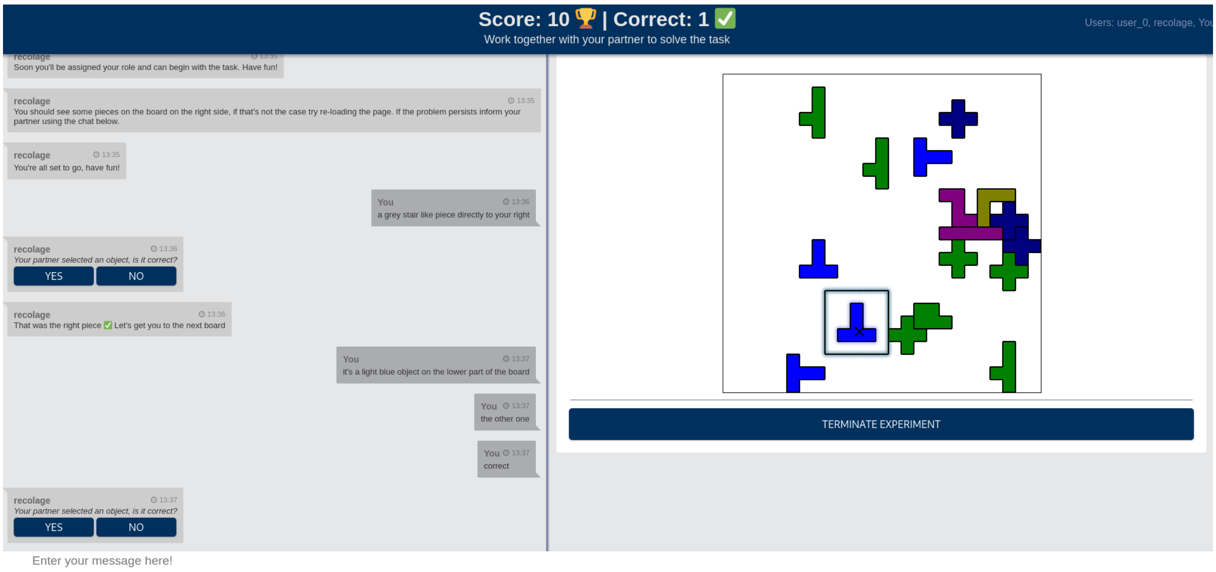
Analogously to work presented by Alacam et al. (2022), participants were shown boards with three different difficulty levels: easy, medium and hard. The complexity of a board is defined by four variables that were used during the process of generation:

- **number of objects:** the total number of objects present on any given board.
- **number of random pieces:** randomly generated pieces are added to the boards to increase variability and prevent generating boards containing only similar pieces.
- **number of similar pieces:** the total amount of pieces on the boards that are grouped based on similar properties. Each piece inside a group shares certain characteristics with other pieces of the same group to add some distractors, thus increasing the complexity of selecting the target object, which is always randomly chosen from one of the grouped pieces.
- **similar pieces per group:** the number of pieces in each group. Grouped pieces share some properties: shape, position, orientation and color. The amount of shared parameters is determined by the difficulty level.

To establish a clear definition of complexity, a pilot study was carried out. Various board settings were explored to identify measurable criteria. The difficulty was measured in terms of speed and number of tokens. The assumption was that a complex task board would require both a higher cognitive load from the participant, and therefore more time to produce it, as well as a higher number of words to describe the target piece.

3.1.1 Pentomino Task Boards

The *Pentomino* boards were generated using the following variables:


Score: 10 🏆 | Correct: 1 ✅
Work together with your partner to solve the task
Users: user_0, recolage, You
TERMINATE EXPERIMENT
YES
NO
Enter your message here!

recolage 13:35
Soon you'll be assigned your role and can begin with the task. Have fun!

recolage 13:35
You should see some pieces on the board on the right side, if that's not the case try re-loading the page. If the problem persists inform your partner using the chat below.

recolage 13:35
You're all set to go, have fun!

You 13:36
a grey stair like piece directly to your right

recolage 13:36
Your partner selected an object, is it correct?

recolage 13:36
That was the right piece ✅ Let's get you to the next board

You 13:37
it's a light blue object on the lower part of the board

You 13:37
the other one

You 13:37
correct

recolage 13:37
Your partner selected an object, is it correct?

Figure 1: *Interface for the instruction giver*

- **shape:** F, I, L, N, P, T, U, V, W, X, Y, Z.
- **color:** red, orange, yellow, green, blue, cyan, purple, brown, gray, pink, olive green, navy blue.
- **position:** top left, top center, top right, left center, center, right center, bottom left, bottom center, bottom right.
- **orientation:** 0, 90, 180, 270.

As mentioned before, similar pieces shared a predefined amount of variables that were fixed within the group. The only exception was made for the position: during the generation of new objects within a group, there is a 50% chance that an object will be assigned a new position instead of the group position to increase variability. The new position is, however, always adjacent to the group position to maintain similarity.

3.2 Variations in the Degree of Interactivity

Four different experiment designs were developed for this data collection to modify the degree of interactivity in the dialogue. The underlying structure of the experiment remains the same across all variations: the instruction giver has to describe the target object to the instruction receiver, who has to select the object on an unlabelled board.

- **No Feedback**: the first variation removes any means of feedback communication between the users. The IG is only allowed to send one single message to the IR, who can then select the described object with a mouse click. After the first message is sent, the IG is not able to write anything else and the players are not notified by the bot whether the IR's selection was correct.

- **Feedback**: while maintaining the same dynamics of the first variation, this variation allows minimal interaction between the users by notifying users about the outcome of each round.

- **Selection Confirmation**: in this variation, interaction between users is enhanced by having the IG confirm the IR's choice once an object has been selected. Upon selecting the wrong piece, the system allowed the IG to send a new description of the target piece. Points are detracted from the total score every time the wrong object is selected.

- **Gripper**: this variation maximizes interactivity between users by not limiting the number of messages that the IG can send. Moreover, object selection by the IG is achieved by moving a gripper on the board instead of using the mouse. The gripper is fully visible for both users at all time allowing the IG to send additional messages correcting or adding new information to ensure the IR moves in the right direction and selects the correct piece.

4 The Data

During task design, the following factors were taken into consideration:

- **Natural:** the IGs were intentionally not provided with any guidance on what constitutes a helpful or accurate description. This decision aimed to force the IGs to generate their own reference expressions independently, without relying on a predetermined pattern.

- **Diverse:** within the same experiments, some variables were modified, hoping that the IG would come up with different descriptions of the target piece, particularly:
  - **Difficulty level:** more complex task boards should require more complex descriptions to uniquely identify the target object.
  - **Degree of interactivity:** different degrees of interactivity between the users and the interface should have a measurable impact on the type of referring expressions generated.

The resulting dataset is a collection of chat logs. 30 participants took part in the experiment and collected a total of 640 data points equally distributed among the four designs and difficulty levels. A single data point is defined as a combination of a *Pentomino* puzzle, the description provided by the instruction giver and the object selected by the instruction receiver. During the entire data collection, the external participants were always assigned the role of the instruction giver, and one experimenter took the role of the instruction receiver.

Every participant was asked to label 20 boards with the exception of two participants who did respectively 79 and 1 to balance the data points across the experiment's variations. Out of a total of 300 pre-generated *Pentomino* boards, 264 were selected randomly by the system at the beginning of every round. On average, each of the 264 boards was selected 2.5 times, with some boards appearing as often as 7 times.

The complete dataset, together with the raw logs and the scripts used to extract and analyze PentoNav are available on Github.

5 Analysis

5.1 Statistical Analysis

In order to run a statistical analysis of the data, the following features were extracted from the dataset:

- batch position: the order of this board within the 20-boards-batch (extracted to measure order effects).
- interactivity: the degree of interactivity.
- complexity: the complexity level of the board.
- accuracy: whether the IR selected the right object after the IG's description. For the interactivity selection confirmation and gripper, the description is marked as corrected if the IG confirmed the correct selection of the IR.
- target: shape of the target object.
- typing lag: how much time (in seconds) the IG took to start typing the referring expression description of the target.
- description lag: how much time (in seconds) the IG took to send the referring expression description of the target.
- response time: how much time (in seconds) the IR took to select an object after receiving the description from the IG.
- number of tokens: number of tokens in the description.
- number of adjectives: number of adjectives used in the description.
- number of adverbs: number of adverbs used in the description.
- number of nouns: number of nouns used in the description.

Before the feature extraction, the descriptions were first normalized with Pyenchant (Pyenchant) and the linguistic features were extracted with *LFTK* (Lee and Lee, 2023). The normalization step consisted of running the spell checker and replacing wrong-spelled words with the first alternative proposed by Pyenchant.

The scope of this analysis is to find out whether the modifications of the experiments had a statistically significant influence on the generated referring expressions. The statistical analysis was carried out using R and the *lme4 package* (lme4).

interactivity	no feedback			feedback			confirm selection			gripper		
complexity	easy	medium	hard	easy	medium	hard	easy	medium	hard	easy	medium	hard
accuracy	85.25	81.13	89.13	93.75	91.67	89.58	100.00	100.00	100.00	93.75	97.92	97.92
lag to typing	6.25	8.79	7.08	7.44	7.89	8.67	5.48	6.44	7.31	4.93	5.49	5.12
lag to description	15.28	18.26	25.83	25.88	29.54	43.46	26.20	30.73	42.40	24.34	22.17	28.02
reaction time	9.08	8.92	10.00	8.73	8.94	11.17	9.17	9.62	13.33	12.41	13.31	16.65
n tokens	6.31	6.91	9.39	8.30	9.27	12.40	10.70	12.81	16.31	13.08	12.17	14.85
n adjectives	1.31	1.13	1.61	1.81	2.08	2.42	1.83	2.27	2.75	1.62	1.38	1.73
n adverbs	0.15	0.17	0.13	0.22	0.15	0.33	0.47	0.50	0.48	0.31	0.23	0.58
n nouns	1.85	2.08	2.54	2.06	2.19	2.96	2.48	2.71	3.52	3.25	2.92	3.38

Table 1: *Mean values of all variables in all levels of complexity and interactivity.*

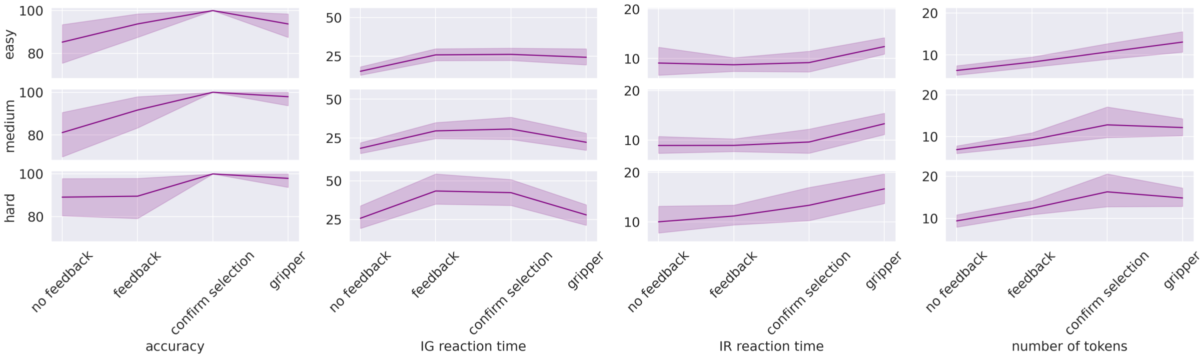


Figure 2: Differences in accuracy, reaction time (IG & IR), and number of tokens per referring expression.

Task Complexity

The results show that the complexity level had a measurable influence on some of the extracted features. The data show that while similar accuracy values can be observed across all three levels, we do see a slight increase in both the time the instruction giver took to both start typing (*lag to typing*) and send the message with the description (*lag to description*). Finally, a substantial increase in reaction time of almost three seconds on the side of the IR can be observed when comparing the easy/medium boards (which have similar values) to the hard scenes.

A look at the linguistic features also indicates that the increasing level of complexity of the board required on average a longer description with an increased number of adjectives and nouns. This initial evidence was also confirmed by training and comparing linear mixed-effects models to fit the data by maximum likelihood with the following parameters:

- fixed effect: complexity level
- random effects: target object, participant, batch position, and interactivity

outcome variable	p-value	χ^2
accuracy	0.7557	0.0968
lag to description	<0.001	41.866
lag to typing	0.03814	4.2989
reaction time	<0.001	11.432
n tokens	<0.001	34.136
n adjectives	<0.001	11.483
n adverbs	0.1123	2.5216
n nouns	<0.001	20.516

Table 2: *Linear mixed effect models: complexity*

All the models were fitted to the data to various outcome variables, which are listed together with the respective *p-values* and χ^2 *values* in *table 2*.

Degree of interactivity

A statistical difference in the data was also found while investigating the effects of the degree of interactivity. The most evident difference can be noted in the accuracy: with increasing levels of interactivity, the accuracy and length of the descriptions also

outcome variable	p-value	χ^2
accuracy	0.04846	3.8939
lag to description	0.7054	0.143
lag to typing	0.05641	3.6399
reaction time	0.02592	4.961
n tokens	0.03131	4.636
n adjectives	0.8686	0.0274
n adverbs	0.07763	3.1138
n nouns	0.06345	3.4448

Table 3: *Linear mixed effect models: interactivity*

outcome variable	p-value	χ^2
accuracy	0.2182	1.5161
lag to description	<0.001	13.18
lag to typing	<0.001	49.852
reaction time	0.4958	0.4638
n tokens	0.9606	0.0024
n adjectives	0.2592	1.2728
n adverbs	0.9336	0.007
n nouns	0.3468	0.8853

Table 4: *Linear mixed effect models: batch position*

increase. Interestingly, the time that the IG needs to start typing decreases while the total time needed to send the description raises from 19.3 seconds in the *no feedback* design to around 32 seconds in both the *feedback* and *selection confirmation* variations to finally fall back to 24.79 seconds in the *gripper* setting. The latter can be tracked down to the fact that in the last setting, the IG was able to send an unlimited number of messages and some users sent multiple shorter messages instead of a longer one, indicating an incremental behavior. While a small increase in the IR's reaction time can be observed when comparing the values of the *no feedback*, *feedback* and *confirm selection* settings, an increase of around 3.5 seconds can be measured in the *gripper* setting. This increase, however, was expected as the gripper is positioned at the center of the board at the beginning of every round and must first be moved on the object that the IR intends to select.

As for the complexity level, linear mixed effect models were also trained to fit the data, and the results are reported in *table 2*. While training the following models, the following parameters were used:

- fixed effect: design
- random effects: target object, participant, batch position, and complexity level

The outcome variables together with the *p-values* and χ^2 *values* are reported in *table 3*.

Batch position

The position of the instance within the batches of 20 boards labelled by the participants also seems to somehow affect the referring expressions with regard to the extracted features. Noteworthy is the effect on the accuracy and the time required by the IG to both start typing and send a message. While for the accuracy, a slight increase can be seen, which indicates that there is a learning effect during the task, this increase only affects the three settings in which the users receive feedback about the piece selected by the IR (feedback, selection confirmation and gripper). The position of the data instance within the batch does not seem to influence the IR's reaction time in any way. With regards to the typing and description lag, on the other hand, a decrease can be measured across all designs.

Linear mixed-effect models were trained with the following parameters:

- fixed effect: batch position
- random effects: target object, participant, interactivity, and complexity

The results are shown in *table 4*.

6 Discussion & Conclusion

In this work, we presented PentoNav: a dataset composed of annotated *Pentomino* puzzles and natural referring expressions generated by the participant to describe one of the objects. The research question postulated at the beginning of this work was whether a manipulation in the degree of interaction between the users and the complexity of the puzzle itself might have an impact on the strategies adopted by the participants to solve the task.

The analysis showed how different degrees of interaction between users, as well as manipulations in task complexity, have a measurable impact on the generated descriptions. Both hypotheses postulated at the beginning of this paper, namely that an increasing level of interaction and puzzle complexity would influence the instruction receiver's accuracy as well as the descriptions generated by the instruction giver, were partially confirmed by the statistical analysis of the data.

One variable that was not considered during planning was the effect of the position of the current data point within the 20 puzzle batches in which the experiment was divided. During the analysis, the position of the data point revealed the learning effect of the participants. This tendency was confirmed by the linear mixed models: while the accuracy increases, the typing and description lag decrease consistently across all designs. This confirms that while progressing through the batch, the participants providing instructions become not only more effective but also faster at generating referring expressions.

6.1 Future work

For the analysis conducted, only a subset of information was extracted from the chat logs. These still contain other information, such as the mouse movements of the instruction receiver on the *Pentomino* board, which can be used to potentially reconstruct the IG's reasoning after receiving a description from the instruction giver. Further insights into the cognitive process of analyzing the description and the board could be offered by the analysis of the instruction receiver's eye movements.

Another interesting application area for this dataset is reinforcement learning. Similarly to the work proposed by Sadler et al. (2023) and Vogel and Jurafsky (2010), an artificial agent can be trained to substitute the instruction receiver and navigate the *Pentomino* board in search of the target piece. Such artificial agents can be deployed online for a subsequent round of data collection, engaging with human participants. The outcome from such agents could be compared to PentoNav to yield valuable insights into the differences between how humans interact with artificial agents versus other human participants.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 423217434 ("RECOLAGE") grant. We would like to thank the anonymous reviewers for their valuable feedback.

References

Özge Alacam, Eugen Ruppert, Amr Rekaby Salama, Tobias Staron, and Wolfgang Menzel. 2020. Eye4Ref: A multimodal eye movement dataset of referentially complex situations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2396–2404, Marseille, France. European Language Resources Association.

Özge Alacam, Eugen Ruppert, Sina Zarrieß, Ganeshan Malhotra, Chris Biemann, and Sina Zarrieß. 2022. Modeling referential gaze in task-oriented settings of varying referential complexity. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 197–210, Online only. Association for Computational Linguistics.

Susan E. Brennan, Katharina S. Schuhmann, and Karla M. Batres. 2013. Entrainment on the move and in the lab: The walking around corpus. In *Cooperative Minds*, Cooperative Minds: Social Interaction and Group Dynamics - Proceedings of the 35th Annual Meeting of the Cognitive Science Society, CogSci 2013, pages 1934–1939. The Cognitive Science Society. Publisher Copyright: © CogSci 2013.All rights reserved.; 35th Annual Meeting of the Cognitive Science Society - Cooperative Minds: Social Interaction and Group Dynamics, CogSci 2013 ; Conference date: 31-07-2013 Through 03-08-2013.

Alasdair Clarke, Micha Elsner, and Hannah Rohde. 2013. Where's wally: The influence of visual salience on referring expression generation. *Frontiers in psychology*, 4:329.

Geneviève de Weck, Anne Salazar Orvig, Stefano Rezzonico, Elise Vinel, and Mélanie Bernasconi. 2019. The impact of the interactional setting on the choice of referring expressions in narratives. *First Language*, 39:014272371983248.

Micha Elsner, Alasdair Clarke, and Hannah Rohde. 2017. Visual complexity and its effects on referring expression generation. *Cogn Sci*, 42 Suppl 4:940–973.

Jana Götze, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen. 2022. The slurk interaction server framework: Better data for better dialog models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4069–4078, Marseille, France. European Language Resources Association.

David Graff, Douglas Reynolds, and Gerald C O'Leary. 1999. Tactical speaker identification speech corpus (tsid).

Meliha Handzic and Graham Low. 2002. The impact of social interaction on performance of decision tasks of varying complexity. *OR Insight*, 15(1):15–22.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 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, Doha, Qatar. Association for Computational Linguistics.

Dimosthenis Kontogiorgos. 2022. *Mutual Understanding in Situated Interactions with Conversational User Interfaces: Theory, Studies, and Computation*. Ph.D. thesis, KTH Royal Institute of Technology.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

lme4. lme4. https://cran.r-project.org/web/packages/lme4/index.html. Accessed: 31/12/2023.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021. Reference and coreference in situated dialogue. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 39–44, Online. Association for Computational Linguistics.

Ramesh Manuvinakurike, David DeVault, and Kallirroi Georgila. 2017. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 331–341, Saarbrücken, Germany. Association for Computational Linguistics.

Hugh Perkins. 2021. Texrel: a green family of datasets for emergent communications on relations. *CoRR*, abs/2105.12804.

Prolific. https://www.prolific.com/. Accessed: 31/12/2023.

Pyenchant. Pyenchant. https://github.com/pyenchant/pyenchant. Accessed: 31/12/2023.

Philipp Sadler, Sherzod Hakimov, and David Schlangen. 2023. Yes, this way! learning to ground referring expressions into actions with intra-episodic feedback from supportive teachers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9228–9239, Toronto, Canada. Association for Computational Linguistics.

Philipp Sadler and David Schlangen. 2023. Pento-DIARef: A diagnostic dataset for learning the incremental algorithm for referring expression generation from examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2106–2122, Dubrovnik, Croatia. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece. Association for Computational Linguistics.

M. Stede. 2012. *Discourse Processing*. Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

University of Edinburgh. 1993. Hcrc map task corpus.

Adam Vogel and Daniel Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 806–814, Uppsala, Sweden. Association for Computational Linguistics.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, Portorož, Slovenia. European Language Resources Association (ELRA).