SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
BACHELOR THESIS
Degree Program: Computer Science (English)

Understanding Pragmatic Reasoning Through Eye Movement Patterns in Reference Games

submitted by
Tymur Mykhalievskyi
Saarbrücken
August 2025

Advisors:
Prof. Dr. Vera Demberg
Computer Science and Computational Linguistics
Saarland University
Saarbrücken, Germany

Dr. John Duff
Computer Science and Computational Linguistics
Saarland University
Saarbrücken, Germany

Reviewer 1: Prof. Dr. Vera Demberg
Reviewer 2: Prof. Dr. Sven Apel

Saarland University
Faculty MI – Mathematics and Computer Science
Department of Computer Science
Campus - Building E1.1
66123 Saarbrücken
Germany

Declarations

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged. I assure that the electronic version is identical in content to the printed version of the thesis.

Tymur Mykhalievskyi

Saarbrücken, August 10, 2025

Acknowledgements

I am extremely grateful to my supervisor, **Dr. John Jack Duff**, for his guidance, patience, responsiveness, and support throughout this project. His thoughtful advice, openness to acknowledging mistakes, and deep expertise in the research topic were invaluable. Without his oversight, I would not have been able to get this far.

Abstract

Reference games are a well-established paradigm for studying pragmatic reasoning, but little is known about how such reasoning is reflected in visual attention. This thesis introduces eye-tracking as a novel method for investigating reference games, using gaze data to gain insight into participants' decision-making processes. By analyzing participants' gaze behavior and self-reported strategies, we examine whether differences in reasoning are tied to how participants attend to and interpret key information. We show that patterns of visual attention reflect meaningful differences in how participants approach the task, even when their overall viewing behavior appears similar. These attentional patterns are predictive of task success and suggest that deeper reasoning may be associated with greater engagement with available messages. The results demonstrate that even low-resolution, webcam-based eye-tracking can reveal valuable information about pragmatic inference, offering a new methodological direction for the study of language and reasoning.

Contents

Introduction

If one says "I am going to Munich this week. My mother lives there.", you will interpret this as meaning they are visiting their mother, even though it is not explicitly stated. This is called an implicature: even without explicitly stating something, one can still convey the information. Human communication is full of such implicit constructions. One reason for this may be to save cognitive effort. What rules do people unconsciously follow during communication to make it more efficient?

In 1975, the British philosopher Paul Grice formulated four types of maxims (Grice, 1975). Maxim of Quantity: Provide as much information as required, but no more. Maxim of Quality: Be truthful, only say that for which you have adequate evidence. Maxim of Relation: be relevant. Maxim of Manner: avoid ambiguity. Returning to the travel example, we can assume a speaker is obeying the maxims. Therefore, the information is relevant and the right amount is provided, so the second sentence about where the mother lives is not just a disconnected fact. Hence, we build an implicature that one is visiting their mother.

One way to study this is through reference games. In these games, participants engage in a collaborative task, often involving the identification or description of objects, where effective communication and reasoning play key roles. Over the years, these reference games have become a popular experimental paradigm to explore how individuals reason about others' intentions and strategies in communication (Frank & Goodman, 2012; Franke & Degen, 2016). A simple example is presented in Figure 1. Imagine someone is talking to you and uses the word "blue" to refer to one of these objects. Which object are they talking about? If you answered blue square, congrats, it is considered the correct solution. It is understandable if this seems confusing at first. If we consider the possibilities of the speaker, the two completely unambiguous messages available to them are "green" and "circle". Hence, if they would have referred to one of the other objects, there are clear messages to do that. Thus, we are left with messages "blue" and "square". Although the message "blue" corresponds to two objects, the blue circle can be referred to unambiguously by using message "circle". Similar logic can be applied to the message "square". Thus, both can be inferred to refer to the blue square. All this reasoning is built upon the Gricean maxims, as we expect from the speaker to be as concise, unambiguous, relevant and truthful as they can be.

On the other hand, one could notice that the reference games are not as intuitive as the traveling example. It is still a limitation that we will have to keep for now. And this study could also shed light on this problem by understanding what exactly people are doing to solve this kind of problems.

In order to deepen the understanding, a formal model was developed, the Rational Speech Act model. It tries to mimic a recursive sequence of reasoning between speaker and listener (Franke & Degen, 2016). This model offers a possible way to derive human-like pragmatic calculations, but much remains untested and unknown about the specific strategies individuals actually employ when they face one of these problems.

This study seeks to expand on prior research by incorporating a novel dimension: tracking participants' eye gaze during reference games. Eye gaze offers valuable insight into how people process information, make decisions, and employ strategies. By capturing where and when participants direct their attention, we can gain a deeper understanding of the cognitive mechanisms at play, including how individuals prioritize certain visual cues and how these cues influence their reasoning strategies. This approach has proven

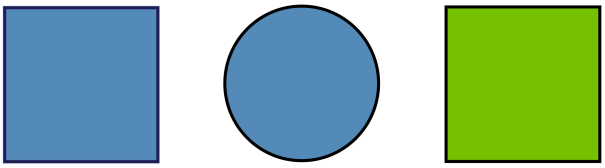


Figure 1: An example of reference game. The same example also appears in Frank and Goodman (2012). A speaker utters "blue", which object are they referring to?

to be a very insightful tool for studying other tasks (Vigneau, Caissie, & Bors, 2006).

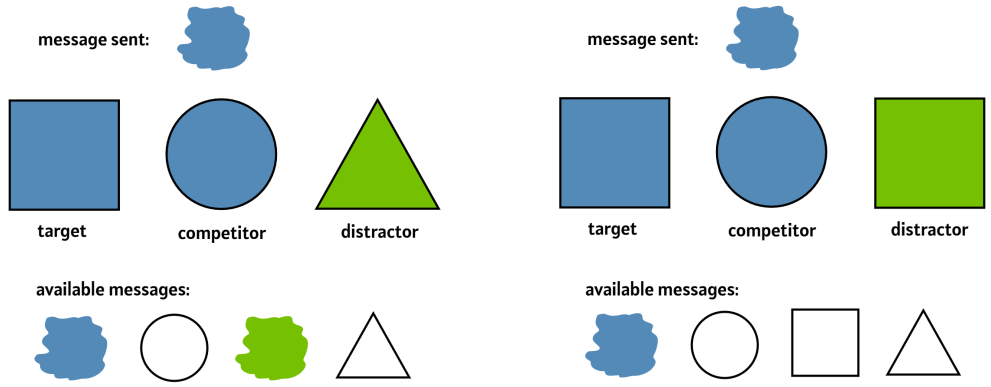
In particular, this study aims to answer the question: *How do gaze patterns correlate with the accuracy and strategies used to solve specific communicative challenges in reference games?* By integrating eye-tracking data with the analysis of reasoning in these games, this paper contributes to a richer understanding of the decision-making processes involved in collaborative communication and problem-solving.

Chapter 1
Related Work

1.1 Reference Games

Although the idea of communication as a signaling game dates back to Lewis (1969), we will focus here on a specific variant known as a *reference game*, as presented in Frank and Goodman (2012). The game is designed to mimic the challenges of everyday communication, as we discussed in the introduction. At first, an instance of the reference game looked as in Figure 1, that is, no information about the available messages is given to the speaker and listener. Later instances, on the other hand, include this information, as illustrated in Figure 1.1. The goal for newer version is the same, that is, to identify which object a speaker is referring to. This change allows for more controlled experiments by increasing the variability in setups. In addition, it should improve clarity, for instance, participants are less likely to question why the speaker would not simply state the object's location instead of describing its properties.

Consider Figure 1.1a more closely. An uttered message is presented on the top. We will denote the object being referred to as a Target, a Competitor is an object that shares the message property with the Target. While a Distractor does not share the sent message property, but could share another property with the Target depending on the difficulty of the trial. Note that, obviously, captions Target, Competitor and Distractor are not available to the participants. The difference between the Simple and Complex trials in Figure 1.1 mainly in how the Distractor is constructed. In particular, in the Simple trial it does not share any properties with the Target, while in the Complex it does. The Simple example Figure 1.1a can be solved without considering the Distractor. That is, one could count the matching messages from the available ones. In this case, it would be 1 for blue square, 2 for blue circle and 2 for green triangle. Hence, the Target is blue square, as "blue" is the only message that could refer to it. This way of solving is not necessarily what people tend to do, but it is one way of interpreting the difference between Simple and Complex trials. Because if you apply the same strategy to the Complex example in Figure 1.1b, both Target and Competitor have two matching messages. On the other hand, if you try to solve these examples yourself, you will probably end up recursively



(a) Simple (b) Complex

Figure 1.1: Two instances of reference games with different difficulties.

reasoning of what the speaker could have said had they had another Target. The Simple and Complex in this case still live up to their names, you will need a more robust recursion in order to solve the Complex one comparing to the Simple one.

1.2 Rational Speech Act Model

Studying this phenomena requires a formalized approach. One such model, called the Rational Speech Act (RSA) was developed. It mimics how a speaker and a listener reason about each other. A detailed explanation can be found in the manuscript by Frank, Emilsson, Peloquin, Goodman, and Potts (2016) as well as in the article by Franke and Degen (2016). We will go through the main ideas on how a listener and a speaker interact with each other. First, consider the matrix M_s in Equation 1.1. Each column is a one-hot encoding of an objects, in other words this matrix encodes which objects match the literal meaning of each message, this matrix is constructed for the Simple example in Figure 1.1a.

$$M_s = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \square & \bigcirc & \triangle \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \tag{1.1}$$

Now let us give an explication for a listener matrix, each row shows conditional probabilities for an objects given a message. Accordingly, a speaker matrix has columns that depict conditional probabilities of messages given an object.

Subsequently we arrive at a literal listener Equation 1.2 and speaker Equation 1.3. Simply put, a literal speaker would output one of the matching messages with equal probability for the given Target. For example, if green triangle is provided for the speaker, they would refer to it by uttering “green” or “triangle” with equal probability. While literal listener would interpret the ambiguous messages with equal probabilities.

$$L_0(M_s) = L(M_s) = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \square & \bigcirc & \triangle \end{array} \\ \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \tag{1.2}$$

$$S_0(M_s) = S(M_s) = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \square & \bigcirc & \triangle \end{array} \\ \begin{bmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0.5 \end{bmatrix} \end{array} \tag{1.3}$$

One could see that such approach would not solve even a Simple trial. However, if there is a completely unambiguous message, the literal listener would be able to correctly identify the Target. These matrices are derived simply by normalizing the message-object matrix along rows (for the listener) or columns (for the speaker). We can keep applying this technique recursively, to find more complex listeners and speakers. That is, a speaker would normalize within columns the matrix previously normalized within rows. In this way we can derive an L_1 listener Equation 1.4, also called a first-order cooperative listener. Note that Frank et al. (2016) and Franke and Degen (2016) apply different strategies to construct L_1 and L_2 listeners, we will stick to the Franke and Degen (2016) variation.

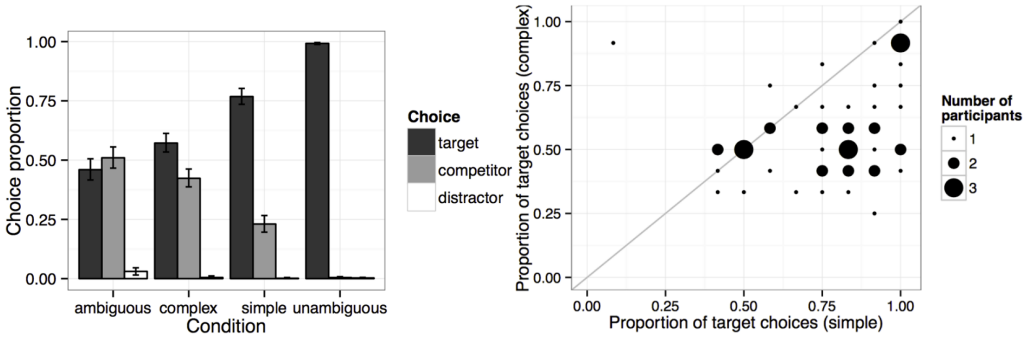
$$L_1(M_s) = L(S(M_s)) = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \square & \bigcirc & \triangle \end{array} \\ \begin{bmatrix} 0.66 & 0.33 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \tag{1.4}$$

The L_1 listener assigns the highest probability to the Target when the message "blue" is used. Repeating this procedure further to get to deeper recursion increases the probability of Target being chosen. In addition, RSA model has a greed parameter α which amplifies the probabilities. $\alpha = \infty$ would result in simply choosing the object with the highest probability. Now let's take a look at the Complex case and see how it differs from the Simple one. The matrix M_c is given in Equation 1.5.

$$M_c = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \square & \bigcirc & \square \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \tag{1.5}$$

Going through the same steps to derive the L_1 listener, we get Equation 1.6.

$$L_1(M_c) = L(S(M_c)) = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \square & \bigcirc & \square \end{array} \\ \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 1 & 0 \\ 0.33 & 0 & 0.66 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \tag{1.6}$$



(a) Proportions of Target, Competitor and Distractor choices in their experiment.

(b) Proportion of Target choices in Simple and Complex conditions by participant.

Figure 1.2: Plots from Franke and Degen (2016).

One important difference is that depth of recursion for the L_1 listener is not enough to assign the highest probability to the Target. Here, the "blue" row has the same probabilities for the Distractor and the Target. Note that in this case the greed parameter α would not be able to help. So instead we consider a deeper level of recursion and introduce an L_2 listener Equation 1.7, also called second-order cooperative listener.

$$L_2(M_c) = L(S(L(M_c))) = \begin{array}{c} \\ \blacksquare \\ \bigcirc \\ \square \\ \triangle \end{array} \begin{array}{c} \begin{array}{ccc} \blacksquare & \bigcirc & \square \end{array} \\ \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0 & 1 & 0 \\ 0.33 & 0 & 0.66 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \tag{1.7}$$

So, as one can see, the L_2 can correctly identify the Target considering the highest probability. Hence, the main point to take from here is that L_1 listener can solve the Simple task, but cannot solve the Complex one, while the L_2 listener is able to solve both.

Further expanding on this, the previous research shows, that the modeled listeners align with the empirical data. In particular, see Figure 1.2 taken from Franke and Degen (2016). Figure 1.2a shows that indeed the difficulty gets harder going from Unambiguous to Simple and further to Complex trials. On the other hand, Figure 1.2b shows that there are roughly 3 clusters present depending on whether one can solve only Simple, both or neither of conditions. This strongly supports the alignment with L_0, L_1 and L_2 listeners. It is important to note that this alignment pertains only to the *accuracy* of the RSA model—not the *strategies* participants actually use.

Now we will proceed further, and make a hypothesis about how people could be solving these problems. A key difference between Simple and Complex trials is the fact that solving Complex trials requires one to consider the Distractor as well due to the matching feature with the Target, while in the Simple one, the Distractor can be ignored completely. This can be demonstrated by the following matrix transformations. If one does not take into account the Distractor the M_s, M_c will instead look as in Equation 1.8 and Equation 1.9 correspondingly.

$$M'_s = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} \blacksquare & \bullet \end{array} \\ \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \end{array} \tag{1.8}$$

$$M'_c = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} \blacksquare & \bullet \end{array} \\ \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \end{array} \tag{1.9}$$

Applying L_1 transformation to the M'_s we get Equation 1.10 which accomplishes the same as in Equation 1.4.

$$L_1(M'_s) = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} \blacksquare & \bullet \end{array} \\ \begin{bmatrix} 0.66 & 0.33 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \end{array} \tag{1.10}$$

On the other hand, neither L_1 nor L_2 can solve the matrix M'_c. In fact, no depth of recursion is helpful in this case as $L_0(M'_c) = L_1(M'_c) = L_2(M'_c)$ (Equation 1.11).

$$L(M'_c) = L(S(M'_c)) = L(S(L(M'_c))) = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} \blacksquare & \bullet \end{array} \\ \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \end{array} \tag{1.11}$$

This leads us to a research question of whether people achieve L_1 accuracy by not considering the Distractor and applying reasoning deeper than one of a literal speaker. Or they include the Distractor in their reasoning but simply lack the depth of recursion therefore failing to solve the Complex trials.

1.3 How Eye Tracking is Useful

Eye tracking has long been used in cognitive science and psychology to study attention, perception, and memory. We will examine at a related field with different kind of tasks, this strategy has shown to be particularly informative and insightful there.

The Raven Progressive Matrices, commonly referred to as the Raven Tests, are a set of nonverbal intelligence tests designed to measure abstract reasoning and problem-solving abilities through pattern recognition and logical inference. Such tests usually contains eight objects arranged in a 3x3 grid with one object missing, along with the set of possible answers displayed below the matrix. Each matrix either has a particular rule it is constructed by or a mix of them. An example is presented in Figure 1.3.

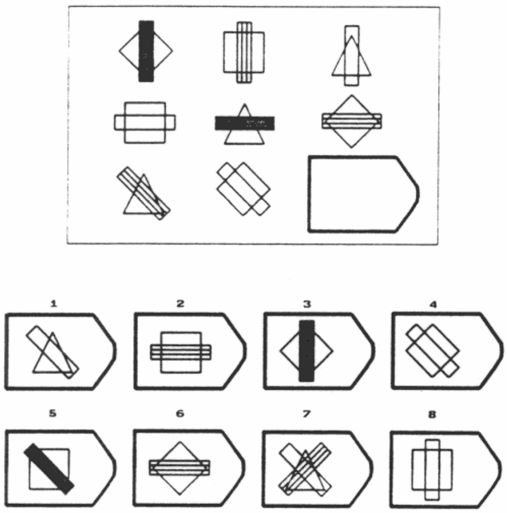


Figure 1.3: An example of Raven Test. The upper part is the matrix, while the bottom is possible answers. The matrix is constructed as follows. The lines orientation is constant within rows. While the shapes and line appearances are obeying the distribution-of-three-values rule. Simply put, same three values are present in each row. The correct answer is 5.

Researches suggest that there are two main strategies for solving the Raven Tests: constructive matching and response elimination (Bethell-Fox, Lohman, & Snow, 1984) and later followed by Vigneau et al. (2006). The former is described as successively identifying rules by which the matrix is constructed, until the answer is fully derived. And the latter means that rather than going through the matrix, one goes over the possible answers and eliminates them one by one, ending up with the correct one in the end. The less efficient of these, response elimination, seemed to be used more frequently by lower-ability subjects on more difficult items.

The two strategies can be identified by the patterns of one's attention, hence, eye gaze. The constructive matching being focused on the matrix and systematically going through rows and columns of it. Carpenter, Just, and Shell (1990) expand on the eye-tracking experiments in this research question by recording eye gaze as well as the verbal comments during the process of solving the tasks. A very detailed sequence of actions is acquired, therefore, giving an insight into how one uses the constructive matching strategy to solve Raven Progressive Matrices. On the other hand, the response elimination involves frequent toggling between the possible answers and the matrix. In order to deepen the understanding in this problem Vigneau et al. (2006) develop a set of features to encode ones attention. Such features include for example Time on Matrix, Time on Alternatives (possible answers) or Number of Toggles between the possible answers and the matrix. The authors report a correlation between the features and the proportion of correct answers. Indeed, the results show statistically significant negative correlation of Time on Alternatives and Number of Toggles with overall score. These findings further support theory about the difference in effectiveness in the two strategies.

One can see, based on these studies, why and how the eye tracking can be useful in the reference games. In our case as discussed in the end of section 1.2 there are multiple ways people could reach L_1 accuracy. Therefore suggesting that the two potential strategies

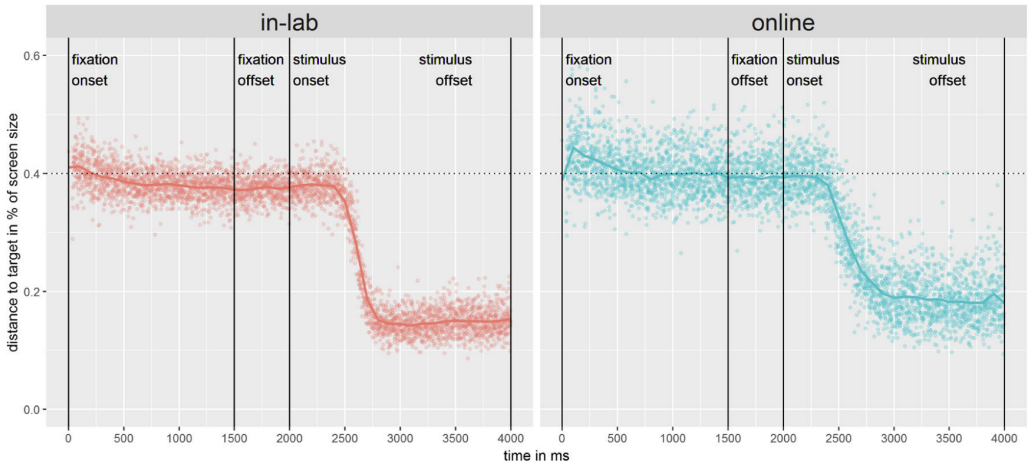


Figure 1.4: Figure from Semmelmann and Weigelt (2018). Fixation task results. Each dot denotes a single recorded data point in distance to a target in percentage of screen size over time.

would be distinguished by the use of Distractor. At the same time, there are still L_0 and L_2 listeners present in the experiment which makes the distinction more difficult. On the other hand, as there is no previous work on eye-tracking reference games, the study is also highly exploratory.

1.4 Out of Lab Eye Tracking

While eye-tracking research has traditionally been conducted in controlled lab settings, there is growing interest in using it in more naturalistic environments. This is important because a reliable and precise method is needed for such experiments. On the other hand, this approach requires participants to be physically present in the laboratory, which makes the experiment far more difficult to conduct in comparison to participants answering a series of questions on their laptops. Therefore, a different approach was chosen. This experiment will incorporate participants' webcams to collect the eye-tracking part of the data. In particular, a library called WebGazer will be used (Papoutsaki et al., 2016). Details about the implementation will be discussed in the following sections.

Although, at first glance, the effectiveness of such approach can be debated, there is work in favor of the method. Starting from the article by Semmelmann and Weigelt (2018), where they take a look into online webcam-based eye tracking comparing it to a respective in-lab experiment. Along with a more fresh research article which also makes this comparison (Wisiecka et al., 2022). Both conclude that while WebGazer is still inferior to the lab equipment in terms of precision, the measurements are reasonably accurate. In particular, taking a look at the results of Semmelmann and Weigelt (2018) shown in Figure 1.4. This figure depicts a particular fixation on the target which was shown after 2000 ms. It takes some time for one to react and for the software to capture the eye movement. Then we observe the saccade in both settings, on average the saccade took 450 ms (750 ms in the online case). The accuracy was 171 px (207 px online), which translates to about 3.94° visual angle in the in-lab setting. In addition, it is visible that the online setting has higher variance.

Taking into account the fact that each problem statement by itself consists of multiple objects located on one page, it is not hard to set them far apart to mitigate the decline in precision. The shorter saccade length is not of the essence in our case. The main goal is to understand the attention profile of the participants. We will talk more in details about the particular challenges and solutions of the setup in the chapter 3.

Chapter 2
Concept

Previous work on individual differences (Franke & Degen, 2016) has focused on variations in underlying pragmatic reasoning tendencies. However, participants with similar underlying pragmatic tendencies may underperform based on the information in the problem that they pay attention to. Take for example the hypothesis we saw in section 1.2, not paying attention to the Distractor would mean that the problem simply becomes unsolvable. Therefore, cognitive abilities such as attention and working memory are most likely not the only features contributing to the performance of participants in these tasks.

This thesis aims to shed light on the realm of where the attention is spent during the pragmatic reasoning problem-solving task. And the eye-tracking data will be used to investigate this.

2.1 Research Questions

2.1.1 Estimating the Posterior

At first, we are interested in predicting the posterior, that is the probability of a participant to solve a problem given their eye gaze features as well as the general information about the trial.

Research Hypothesis 1

As we discussed in section 1.2, the Distractor in the hypothesis is a crucial part of the problem. It is possible that one reason why a participant might struggle with Complex trials is because they do not consider the Distractor to be relevant. Therefore, it is important to understand how participants interact with the Distractor. The first research hypothesis is: Proportional time on Distractor is positively associated with accuracy on Complex trials.

Research Hypothesis 2

The second research question is about the messages that are available to the participants. The second research hypothesis is: Proportional time on available messages is positively associated with accuracy on Simple trials. This hypothesis is based on the idea that while the Unambiguous trials can be solved without considering the available messages, the Simple ones require one to consider the available messages.

In addition, the Complex trials can be solved without looking at the available messages at all. For example, in the trial in Figure 1.1b. Taking into account the available messages the following correspondence can be built: "circle" – Competitor (blue circle), "square" – Distractor (green square) and "blue" – Target (blue square). However, if one does not consider the available messages, they can build a similar correspondence: "circle" – Competitor (blue circle), "green" – Distractor(green square) and "blue" or "square" – Target (blue square). In both cases the message "blue" corresponds to the Target.

2.1.2 Estimating the Likelihood

Because Complex trials require one to consider the Distractor, while Simple trials do not, and because Simple trials require one to consider the available messages, while Complex trials do not, it is possible that accurate participants have different attention profiles in Simple and Complex trials. Therefore, we would like to do a slightly alternative analysis, estimating the likelihood directly instead of trying to estimate the posterior. In order to realize this, we will only take the correctly solved trials into account and predict the probability of a each fixation (gaze point) to be on the area of interest.

Research Hypothesis 3

The third research hypothesis is: on correctly solved trials, Complex trial condition is positively associated with the probability of fixation (gaze point) being on the Distractor.

Research Hypothesis 4

The fourth research hypothesis is: on correctly solved trials, Simple trial condition is positively associated with the probability of fixation (gaze point) being on the available messages.

Chapter 3
Method

3.1 Data Collection

The experiment was hosted through an open-source crowdsourcing client/system server system LingoTurk (Pusse, Sayeed, & Demberg, 2016). The participants were recruited via Prolific (Palan & Schitter, 2018), a subject pool for online experiments. The following criteria were used to filter the participants: native English speaker located in the UK, age in range 18-40, minimal approval rate of 95%, number of previous submissions must be at least 20 studies, participants cannot have taken part in any of the related studies our group has conducted before. The estimated length of the experiment was 22 minutes, with the median completion time being 19 minutes and 30 seconds. The participants were paid 3.89 pounds which is equivalent to minimal hour wage in Germany. The participants were paid only in case of successful completion of the full experiment. The experiment was conducted in a web browser and the participants were asked to use a computer with functional webcam. The project was approved by the Ethical Review Board of the Faculty of Mathematics and Computer Science at Saarland University.

3.1.1 Reference Games

There were 3 conditions of the reference games used in this experiment: Simple, Complex and Unambiguous. In the section 1.2 we mainly talked about the Simple and Complex conditions. The Unambiguous condition serves two main purposes. First of all, it acts as a sort of filler, so that participants do not get used to the same type of problems. Second of all, it acts as a control check, that is, participants who do not reach 75% accuracy on the Unambiguous trials are excluded from the analysis.

The trials were generated using 3 colors: red, green and blue, and 3 shapes: square, circle and triangle. The code used to generate trials can be found under the folder “trials” (Mykhalievskyi, 2025). For each condition every unique sent message was repeated exactly twice. Which results in 12 trials of every condition. Another restriction we included was that available messages always have 2 shapes and 2 colors. Furthermore

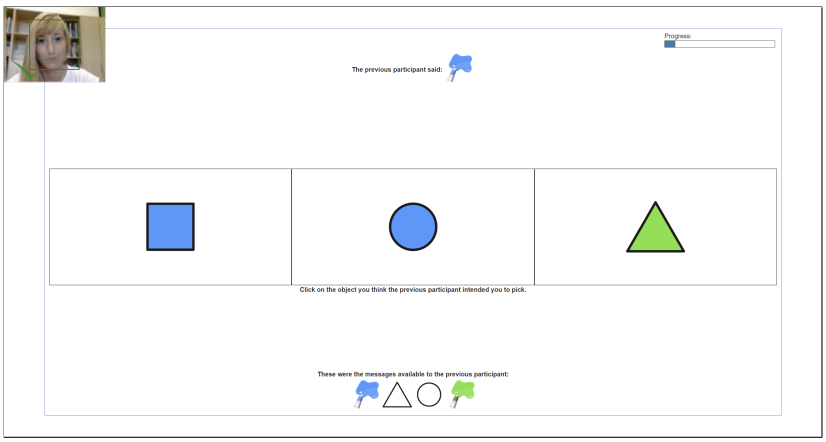


Figure 3.1: Example trial

one Simple and one Complex trial was picked to be repeated once again in the very end of the experiment. These were, so called, strategy trials where we would ask the participants to explain their reasoning behind the choice of the object, a similar approach was incorporated in the study by Mayn and Demberg (2023). All trials were randomly shuffled before the experiment, except for the strategy trials which were always at the end of the experiment. The full list of trials is shown in in Table B.1.

In addition, a feedback system was incorporated in the trials. After every answer a pop up at the top of the screen would show up for 1200 ms. It would only show "Correct" or "Incorrect", based on whether the participant chose the Target or not. No other feedback was given during the experiment. The strategy trials did not include this feature at all.

Before the main part of the experiment with 36 + 2 trials, the participants were asked to do a speaker's job in reference games with 3 Unambiguous trials and 1 completely ambiguous one. The participants were asked to describe the object in the ambiguous trial in a way that the listener would be able to pick the correct object. Later they were told that a previous participant had already done the speaker's job and they were to do the listener's job, which was the main part of the experiment with 36 + 2 trials. An example trial can be seen in Figure 3.1. Depending on the zoom and the resolution of the page, the sizes and absolute positions of the images would vary a lot. Hence, the absolute positions in pixels as well as sizes of the images were saved during the experiment.

3.1.2 Eye Tracking

The eye tracking was done via library WebGazer (Papoutsaki et al., 2016). The library was used to track participants' gaze on the screen. The library is implemented as a linear regression model to predict x and y coordinates based on encoded gaze and facial features. There was a calibration in the beginning of the experiment after the practice trials where participants did the speaker's job, this allowed to put the calibration as close to the main experiment as possible. The calibration was adapted from the one used in the demo of WebGazer (Papoutsaki et al., 2016). During the calibration participants are asked to click on the points, every click is used to adjust the model parameters to the particular setup. The calibration works based on the assumption that people would look where they click. We included 11 points instead of 9 as in the demo, the additional points were

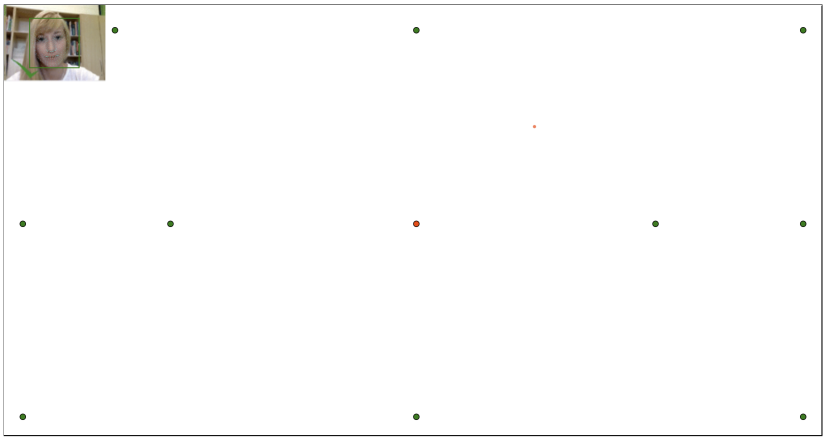


Figure 3.2: Calibration setup

put on the objects' places. Each point had to be clicked 5 times. The setup can be seen in Figure 3.2.In addition, the calibration accuracy assessment in the end was done not with 1 point but 3: middle, left and right. Where left and right again correspond to the positions of the main objects on the screen. Furthermore, an in-between trial calibration was incorporated to ensure that the calibration was still accurate. This drastically improves the performance of the model as during the experiment participants would make small or not adjustments which would make the initial calibration not as accurate as they progress through the experiment. However, including the in-between calibration solves this problem as right before the trial, participants click and by assumption look at the places where the areas of interest are located making the model adjust to the current setup. The model only saves last 50 clicks and replaces the oldest ones one by one after more than 50 were done. During the in-between calibration the number of points were reduced to 5, each point corresponding to one of the areas of interest. No accuracy assessment was performed during the in-between trial calibration. The setup can be seen in Figure 3.3. In order to make the first fixation after the in-between calibration more predictable, the top point, that corresponds to the sent message, was not available until all the other points were clicked on. This way participants would enter the trial looking at the sent message which is we expected to be the first point of interest. The sent message gives crucial information about the trial without which it is impossible to solve the task.

In order to successfully pass the calibration assessment, the participants must reach at least 65% accuracy on each of the three calibration points. However, during the testing phase we noticed that the calibration assessment was too strict and difficult to pass. Therefore, it was determined that the left and right points should be made easier to calibrate via adjusting the calculation of the accuracy. A weighted accuracy calibration procedure was implemented. While the distance for the accuracy of the middle point was calculated via euclidean distance, the left and right points were calculated via the following formula: $\sqrt{(w_x \cdot (calib\_point\_x - gaze\_x))^2 + (w_y \cdot (calib\_point\_y - gaze\_y))^2}$. Where w_x and w_y are the coefficients that were adjusted during the testing phase. The final values were $w_x = 1$ and $w_y = 0.5$. The values were chosen based on the fact that left and right objects do not have any other objects on the vertical axes this can be seen in Figure 3.1. Therefore a slightly inaccurate result on the vertical axes would be relatively easy to correct during the analysis.

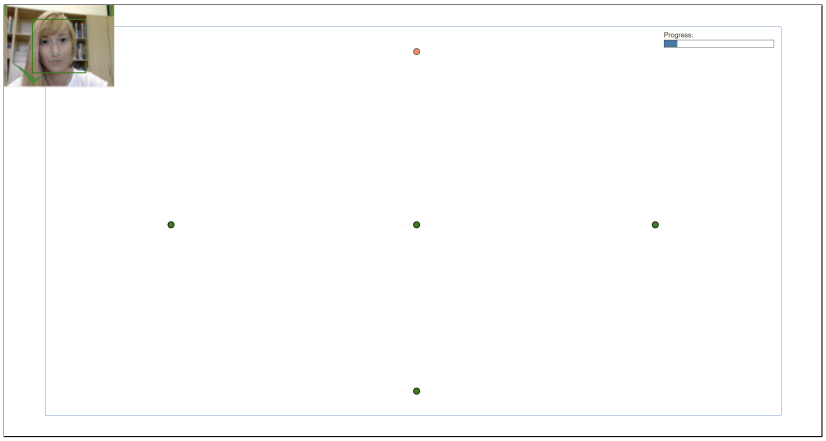


Figure 3.3: In-between trial calibration

Figure 3.4: Calibration instructions taken from Semmelmann and Weigelt (2018)

The images were located on the screen as far as possible to reduce the errors as much as possible. For the same reason, the available messages were kept as a single block instead of being spread further apart.

Furthermore, a performance issue arised during the pilot phase. The issues was that the eye tracking became very slow and laggy towards the the second half of the experiment. The more clicks were made, the worse the performance became. Due to a drastic drop in sampling rate and increase in response time, the pilot data was clearly unacceptable. The issue was resolved by reducing the DataWindow size from 700 to 50 in source code and recreating the WebGazer source file afterwards. The issue was resolved and the performance was stable throughout the experiment.

The calibration is highly dependent on the setup. Hence, the following pieces of advice were given to the participants to increase chances of successful calibration: keep the laptop on charging during the whole experiment (to make sure the eye tracking does not suffer from battery saving features); choose a quiet, well-lit room with minimal distractions and use a stable chair; place your laptop on a stable surface, screen directly in front of you. Later during the calibration, ensure the webcam is centered with your face; keep your head as still as possible during the experiment; make this window full screen size if not already. In addition, right before the calibration the participants were shown their webcam feed to make sure they are centered and the lighting is good as well as some visual advice on how to improve the calibration. The visuals can be seen in Figure 3.4. They were taken from a different experiment (Semmelmann & Weigelt, 2018) that was conducted with use of WebGazer (Papoutsaki et al., 2016).

Even with all the optimizations in place, the calibration procedure was still relatively difficult to pass. The participants were not restricted in number of attempts to calibrate. However, the participants were informed that the approval is only possible after

successfully completing the calibration and the rest of the experiment.

3.1.3 Consent

The participants were informed about the eye-tracking procedure at the beginning of the experiment. They were informed that no video of them will be stored at any point of time. The participants were informed that they can stop the experiment at any point by exiting the website and none of their data will be saved. The project and the consent formulation was approved by the Ethical Review Board of the Faculty of Mathematics and Computer Science at Saarland University. The exact formulation was:

> This experiment is being conducted as part of ongoing research at Saarland University. If you have any questions or comments about the study, please contact us. You must be at least 18 years old to participate. Your participation in this research is voluntary. There are no risks or benefits to participating in this study. You may decline to answer any or all of the following questions. You may decline further participation, at any time, without adverse consequences. Part of the data collections involves using your webcam to estimate your eye gaze. No video or audio is stored at any point during the experiment. We only use the video to store the estimated position of your eye gaze, as well as estimated size of your pupil. All data will be anonymized prior to analysis.
>
> If you agree to participate, please read the below instructions before proceeding.

3.2 Analysis

3.2.1 Features

Considering the related eye-tracking study conducted by Vigneau et al. (2006), the analysis will be around the defined eye-tracking features. Vigneau et al. (2006) defined mainly 5 types of features: absolute time on an area of interest, proportional time on an area of interest, toggling (between the Raven's matrix and the available messages), item latency (time to complete the trial) and latency to first toggle and matrix distribution index (how equally the attention was spread across the matrix items). In this study on the other hand, we are mainly focused on the proportional time on the areas of interest. The decision was made for multiple reasons. First of all the sampling rate of WebGazer is highly dependent on the participants' computer, making the absolute time on the areas of interest and latency to first fixation not comparable between participants. Second of all, we do not have such a strong hypothesis about what exactly people do during the task solving process as Vigneau et al. (2006) had. We are rather interested in the general profile of attention, hence, no toggling features were included in the analysis.

In addition to the eye-tracking features, we included some general features about the trial. The features are: trial number, condition (Simple, Complex or Unambiguous), type of sent message feature (shape or color), correct (whether the trial was solved correctly or not) and Target position (left, center or right). The condition and correct features play crucial role in the analysis. While the other features such as trial number, type of sent message and Target position were included due to the fact that they were shown

to sometimes have a significant effect on the participants' performance in the previous studies (Mayn & Demberg, 2023; Mayn, Loy, & Demberg, 2025).

3.2.2 Data Preprocessing

The data preprocessing was done in Python using the pandas library (Reback et al., 2020). Firstly, due to how LingoTurk (Pusse et al., 2016) is implemented, we had to parse the data from string into the dictionary and create a data frame from it. As was mentioned before, the WebGazer (Papoutsaki et al., 2016) library was used to track the eye gaze. The WebGazer predicts the x and y coordinates of the current gaze point on the screen. Hence, we had to define a certain boundaries for the areas of interest in order to assign the predicted gaze points to them. Taking into account the weighted calibration described in subsection 3.1.2, which was implemented for the left and right objects we had to make the margin a little larger for the left and right objects comparing to the center one. We settled on defining four point polygons for each of the interest areas.

The polygons were defined using the following variables:

- x_{12} – x coordinate equally distant between the left and the center images of the objects.
- x_{23} – x coordinate equally distant between the center and the right images of the objects.
- y_{12} – y coordinate equally distant between the sent message image and the center object image.
- y_{23} – y coordinate equally distant between the center object image and the available messages images.

It is worth to note that the values of the variables can be easily computed using the coordinates of left top corner of an image and its width and height which were saved during the experiment. Finally, the polygons were defined via four points as follows:

- sent message – $((x_{23} + \frac{x_{12}}{2}, 0), (\frac{x_{12}}{2}, 0), (x_{12}, y_{12}), (x_{23}, y_{12}))$
- left object – $((0, \frac{y_{12}}{2}), (x_{12}, y_{12}), (x_{12}, y_{23}), (0, y_{23} + \frac{y_{12}}{2}))$
- center object – $((x_{12}, y_{12}), (x_{23}, y_{12}), (x_{23}, y_{23}), (x_{12}, y_{23}))$
- right object – $((x_{23}, y_{12}), (x_{12} + x_{23}, \frac{y_{12}}{2}), (x_{12} + x_{23}, y_{23} + \frac{y_{12}}{2}), (x_{23}, y_{23}))$
- available messages – $((x_{12}, y_{23}), (x_{23}, y_{23}), (x_{23} + \frac{x_{12}}{2}, y_{12} + y_{23}), (\frac{x_{12}}{2}, y_{12} + y_{23}))$

The visualization is shown in Figure 3.5.

3.2.3 Eye Tracking Features

Based on whether a point was inside a certain polygon or none of them it was assigned to the corresponding area of interest or to the non-area-of-interest category. Each point was not represented by the time between the current and the previous point as the rate of sampling was not the same across participants. In order to mitigate this issue a fixation detection algorithm was tried out. The algorithm is implemented in R (von der Malsburg,

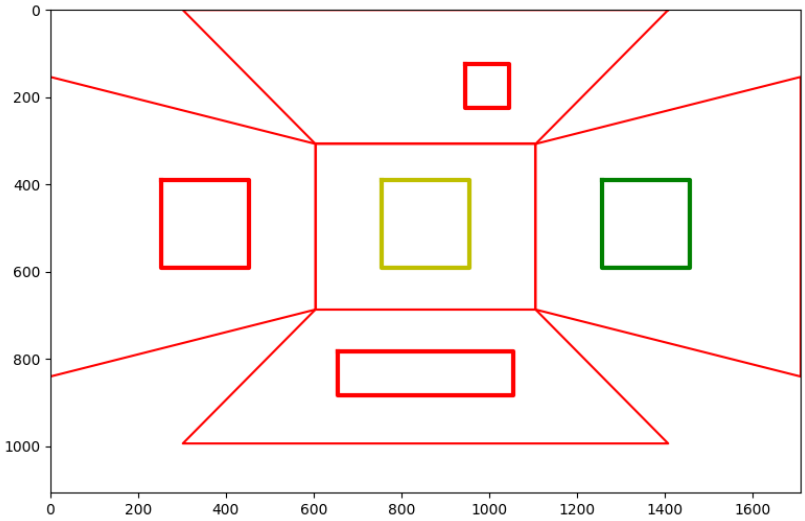


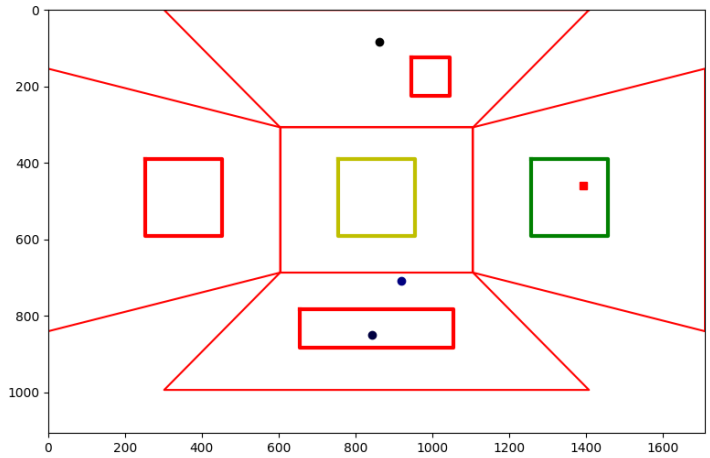
Figure 3.5: Polygons for the areas of interest. The inner rectangles show the exact positions of the images on the participant's screen. The outer rectangles show the polygons used to define the areas of interest.

2015). The algorithm is based on the fact that due to a fixation present some points would be close to each other while the rest would be far away. The algorithm states to work well even for a low quality data with sampling rate less than 100 Hz. However, in our case an average sampling rate amounted to 18 Hz. Even with the lowest tolerance, the algorithm was not able to detect all fixations accurately due to the low sampling rate. An example of fixation detection and the original scanpath can be seen in Figure 3.6. Here, the algorithm correctly detected 4 fixations, they are: on the sent message, twice on the bank of available messages and the last one being on the right object. While the second fixations of the available messages is debatable, clearly the algorithm was not able to detect the quick glance on the middle object before the last fixation on the right object. Clearly, the sampling rate was too low in order to determine the fixations accurately. Hence, the algorithm was not used in the analysis.

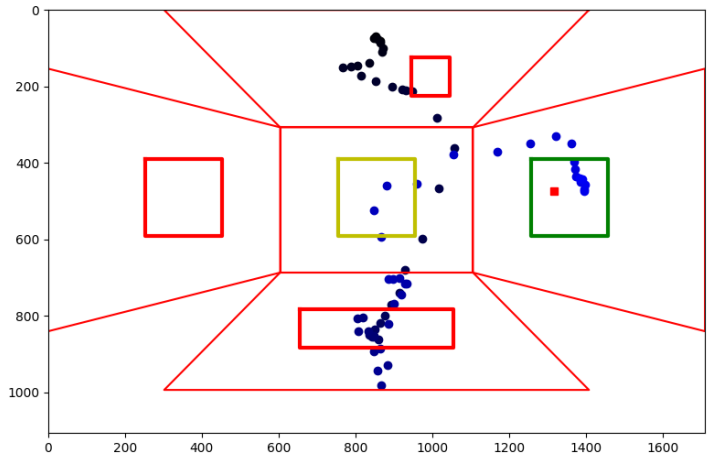
Instead of using the algorithm, we decided to use the raw scanpath and derive the features from it. We assign each predicted gaze point to the corresponding area of interest. Then count the amount of them in each area of interest. This way we derive absolute time on area of interest. Again, it is worth mentioning that we did not use the actual timings of the predictions because the sampling rate varied a lot across the participants. Hence, we used the amount of points in each area of interest as a proxy for the time spent on it. Furthermore, we calculated the proportional time on area of interest as the amount of points in the area of interest divided by the total amount of points in the trial, including the ones that missed the areas of interest. The proportional time on area of interest is the main feature used in this thesis. An example of calculated features for the scanpath in Figure 3.7 can be seen in Figure 3.1.

3.2.4 Pairwise Correlations

Similarly to Vigneau et al. (2006), we will report the pairwise correlations of eye-tracking features and the accuracy based on condition. This should give us some general idea



(a) Fixations



(b) Scanpath

Figure 3.6: Comparison of the raw scanpath (b) and the detected fixations (a). The hue of the points corresponds to the order of the points. Black being the first point and the pure point being the last. In addition, the very last point is marked with a red square.

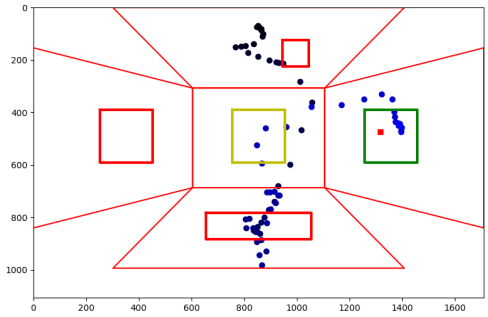


Figure 3.7: Scanpath of a Simple trial, and the layout of areas of interest. The colors of the main objects are used as follows: green – Target, yellow – Competitor and red – Distractor.

Feature	Value
TimeOnSentMsg	18
TimeOnAvailableMsgs	27
TimeOnTrgt	14
TimeOnDist	0
TimeOnComp	9
TimeOnNonAOI	0
PropTimeOnSentMsg	0.26
PropTimeOnAvailableMsgs	0.4
PropTimeOnTrgt	0.21
PropTimeOnDist	0.0
PropTimeOnComp	0.13
PropTimeOnNonAOI	0.0

Table 3.1: Features and their corresponding values. The features starting from "Time" corresponds to amount of points in the area of interest.

about what features are correlated with the accuracy. In addition, the correlation will test some of our hypothesis.

3.2.5 Mixed Effects Logistic Regressions

Throughout our modeling, we assume that the data are independent and identically distributed (i.i.d.), meaning that each trial is treated as an observation drawn from the same distribution and independent of other trials. The main analysis will be done using mixed effects logistic regressions. The models are designed to correspond to our hypothesis presented in section 2.1. For the hypothesis 1 and 2, the following model was preregistered (Mykhalievskyi, Duff, & Demberg, 2025):

```
Correct ~ Condition + TrgtPos + Trial +
PropTimeOnTrgt + PropTimeOnComp + PropTimeOnDist +
PropTimeOnSentMsg + PropTimeOnAvailableMsgs +
PropTimeOnNonAOI + MsgType +
(1 + Condition + TrgtPos + Trial + PropTimeOnTrgt +
PropTimeOnComp + PropTimeOnDist + PropTimeOnSentMsg +
PropTimeOnAvailableMsgs + MsgType | Subject)
```

However, we quickly realized that the model has a few issues. First of all, the model is missing interaction between the different conditions and the eye-tracking features. Therefore, the interaction terms were added to the model. The second issue is that the feature "PropTimeOnNonAOI" can be derived from the other eye-tracking features making it redundant and unnecessary to include in the model, hence, it was removed. In addition, we added the "AnswerTime" feature to the model. The "AnswerTime" feature is the time between the start of the trial and the moment when the participant clicked on the answer. It was added to account for the fact that some participants might have been faster than others. The final starting model looks as follows:

```
Correct ~ Condition + TrgtPos + Trial +
PropTimeOnTrgt + PropTimeOnComp + PropTimeOnDist +
PropTimeOnSentMsg + PropTimeOnAvailableMsgs + MsgType +
AnswerTime +
Condition:PropTimeOnTrgt + Condition:PropTimeOnComp +
Condition:PropTimeOnDist + Condition:PropTimeOnSentMsg +
Condition:PropTimeOnAvailableMsgs + Condition:AnswerTime +
(1 + Condition + TrgtPos + Trial + PropTimeOnTrgt +
PropTimeOnComp + PropTimeOnDist + PropTimeOnSentMsg +
PropTimeOnAvailableMsgs + MsgType + AnswerTime | Subject)
```

Through this model we will be able to test the first two hypothesis. Mainly whether the Distractor is positively associated with the accuracy on the Complex trials. And, for the second hypothesis, whether the available messages become important on the Simple trial comparing to Unambiguous and maybe Complex as well. The model will be fit on the trial level and predicts whether the answer was correct or not. The "Trial" and "AnswerTime" features were rescaled to be between -1 and 1, in order to match the scale of other features. All the eye-tracking features were centered to have mean 0 but not scaled. This approach was chosen because it would make the interpretability of the models easier.

As for the third and fourth hypothesis, each of them will be tested using a separate model. These models treat each fixation as an independent observation. Accordingly, we assume the gaze data to be i.i.d., meaning that each fixation is considered to be drawn independently from the same probability distribution. Both of them predict whether a gaze point was on a certain area of interest. Hence, the data to fit them is made on the gaze point level. Only the gaze points from the correctly solved trials were added to the data to match the hypothesis. The third hypothesis will be tested using the following model:

```
OnDist ~ Condition + Trial + MsgType + TrgtPos +
(1 + Condition + Trial + MsgType + TrgtPos | Subject)
```

It predicts whether a gaze point was on the Distractor or not based on the general information about the trial.

The fourth hypothesis will be tested using the following model:

```
OnAvMsgs ~ Condition + Trial + MsgType + TrgtPos +
(1 + Condition + Trial + MsgType + TrgtPos | Subject)
```

It predicts whether a gaze point was on the bank of available messages or not based on the general information about the trial.

Chapter 4
Results

4.1 General Information

In total there were 120 participants in the study who made a submission according to Prolific. 12 were removed due to technical issues with the submissions, either their submissions were not received due to technical issues or they actually did not fully complete the experiment. Another 3 were removed due to having missing data in some of the trials and another 4 were removed due to having accuracy on Unambiguous trials below 75% according to our preregistration (Mykhalievskyi et al., 2025). This left us with 101 participants for the analysis.

It is worth noting that due to the nature of the experiment, the calibration was challenging to pass, making the experiment difficult to complete. 120 participants completed the experiment, however, there were around 150 submission attempts that were not completed. This mainly happened because of the calibration difficulties, according to feedback from some participants. This information gives us an idea about the completion rate of the experiment, which amounted to around 40%. However, this allowed us to collect a large amount of good quality data, which is the most important aspect of the experiment.

Due to the complexity of the mixed effects models, convergence issues arose. Therefore, the random effects were removed one by one from the models based on the least variance among the random effects. The process was repeated until the models converged. In the case of predicting accuracy, the random effects were removed from the model entirely.

The general information about the accuracy on the participant level can be seen in Figure 4.1 and Figure 4.2. The Figure 4.1 demonstrates that L_2 reasoning takes the largest portion of the participants. This is probably the case due to the included feedback in the trials, which greatly improves the probability of a participant learning a strategy to solve the trials from all 3 conditions. However, generally the data follows similar pattern to the one reported in Franke and Degen (2016), similar scatter plot can be seen in Figure 1.2b. It is worth to note that a few participants achieved extremely low accuracy on Simple trials, while performing relatively better on Complex trials. The cause of this is unclear.

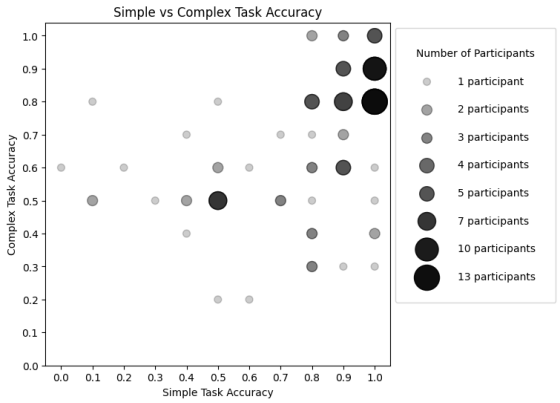


Figure 4.1: Scatter plot of accuracy.

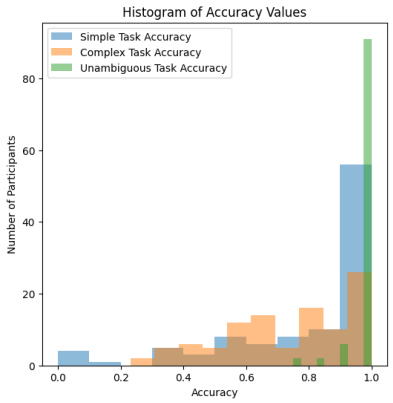


Figure 4.2: Histogram of accuracy.

The histogram in Figure 4.2 shows that the general pattern of people easily solving the Unambiguous trials, struggling more in the Simple condition and finding the Complex condition the most difficult to solve.

4.2 Pairwise Correlations

Feature	Mean (SD)	Accuracy	Mean Answer Time
PropTimeOnSentMsg	0.29 (0.09)	-0.33 ***	-0.57 ***
PropTimeOnAvailableMsgs	0.11 (0.07)	0.35 ***	0.38 ***
PropTimeOnTrgt	0.24 (0.06)	0.35 ***	0.18
PropTimeOnDist	0.16 (0.06)	-0.03	0.29 **
PropTimeOnComp	0.18 (0.06)	-0.24 *	-0.07
PropTimeOnNonAOI	0.02 (0.01)	-0.02	0.04
RateTogglingAvailableMsgs	0.04 (0.02)	0.41 ***	0.44 ***
NumTogglesAvailableMsgs	4.39 (3.23)	0.18	0.65 ***
MeanAnswerTime (ms)	5188 (2586)	0.08	—
Accuracy	0.8 (0.25)	—	—

Table 4.1: Simple condition. Correlation table showing the relationships between features, accuracy, and mean answer time. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The pairwise correlations for the Simple and Complex conditions can be seen in Table 4.1 and Table 4.2 correspondingly. The correlations between the eye-tracking features were excluded due to the way they were defined. Almost all of them have negative correlation because the features are proportional and increase in one of them means a decrease in some of the others. However, there was one exception to this general trend, it is a positive significant correlation of 0.2 between the `PropTimeOnDist` and `PropTimeOnNonAOI` on Complex condition. This indicates that on Complex trials, the more time a participant spends on the Distractor, the more time they spend outside of the AOI. Generally the Non-AOI feature is extremely low comparing to other eye-tracking features. The only reasonable explanation is that due to the increase in difficulty in the Complex trials, the participants are more likely to have a more of an equally distributed attention across the areas of interest.

Looking at the correlations with accuracy, there are some significant effects. For both conditions, the two less interesting ones are the correlations with `PropTimeOnTrgt` and `PropTimeOnComp`. In the former case, there is a significant positive correlation between `PropTimeOnTrgt` and `Accuracy`. The more participant look at the Target, the more likely they are to answer correctly. This is not surprising as the Target is the correct answer. A similar logic can be applied to the later, there is a significant negative correlation between `PropTimeOnComp` and `Accuracy`. The more time a participant spends looking at the Competitor, the more likely they are to choose it, therefore, answering incorrectly. In addition, we visually inspected the scanpaths of some participants. The effect might be coming from the last look at the object of choice, therefore, tipping the proportional time in favor of the chosen object and, hence, correlating the proportional time to the answer.

Feature	**Mean (SD)**	**Accuracy**	**Mean Answer Time**
PropTimeOnSentMsg	0.26 (0.09)	-0.35 ***	-0.61 ***
PropTimeOnAvailableMsgs	0.13 (0.07)	0.22 *	0.51 ***
PropTimeOnTrgt	0.23 (0.07)	0.56 ***	0.02
PropTimeOnDist	0.15 (0.06)	-0.08	0.19
PropTimeOnComp	0.21 (0.06)	-0.26 **	0.11
PropTimeOnNonAOI	0.02 (0.02)	-0.01	0.01
RateTogglingAvailableMsgs	0.04 (0.02)	0.25 *	0.45 ***
NumTogglesAvailableMsgs	5.19 (3.69)	0.23 *	0.78 ***
MeanAnswerTime (ms)	6565 (4022)	0.18	—
Accuracy	0.7 (0.21)	—	—

Table 4.2: Complex condition. Correlation table showing the relationships between features, accuracy, and mean answer time. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Furthermore, the `PropTimeOnSentMsg` has a significant negative correlation with accuracy in both conditions. This indicates that the more time a participant spends looking at the sent message, the less likely they are to answer correctly. This can be explained by the fact that the sent message is important in the trial, however, spending more time looking at it would mean that less time is spent looking at the other crucial areas of interest. And a better accuracy is achieved probably by taking a quick look at the sent message and then focusing on the other areas of interest.

As for the `PropTimeOnAvailableMsgs`, it has a significant positive correlation with accuracy in both conditions. This indicates that the more time a participant spends looking at the available messages, the more likely they are to answer correctly. This is not surprising as the available messages are extremely important to solve the Simple trials. It is worth noting that the effect is much stronger in the Simple condition, where the correlation is 0.35, comparing to the Complex condition, where the correlation is 0.22. This partially aligns with the second hypothesis described in subsubsection 2.1.1. The effect is significant in both conditions. While we anticipated that the Simple condition would benefit from the available messages more, the Complex condition also benefits from it. This is probably due to the fact that the available messages are still important for the participants in the Complex trials.

Both `RateTogglingAvailableMsgs` and `NumTogglesAvailableMsgs` show significant positive correlations with accuracy in Complex condition. However, in simple condition, only the `RateTogglingAvailableMsgs` shows a significant positive corre-

lation with accuracy. This indicates that the raw amount of toggles is not important by itself, but only shows up as significant when combined with answer time. Potentially indicating that a certain successful strategy is encoded in the high rate of toggling for Simple trials.

Now, looking at the column of `MeanAnswerTime`. We did not make any hypothesis about how answer time would relate to the eye-tracking features, however, it can still give us some general information about participant-level patterns. The `PropTimeOnSentMsg` has a significant negative correlation with the answer time. The reason for this is probably related to the interpretation of the correlation with the accuracy. Participants might adapt a suboptimal strategy where they would not look at the available messages which works perfectly for the Unambiguous trials, however, completely fails on the Simple trials. That is why the proportional increase in time spent looking at the sent message indicates a quick answer and low accuracy.

The `PropTimeOnAvailableMsgs` has a significant positive correlation with the answer time. This indicates that the more time a participant spends looking at the available messages, the more time they would spend on the trial in general. This is not surprising as proportionally spending more time looking at the available messages indicates more reasoning involved in the problem solving process, which leads to more time spent on the trial. Especially taking into account the fact that the available messages has second to lowest average time spent on it in both conditions.

`NumTogglesAvailableMsgs` have significant positive correlations with the answer time in both conditions. This is expected as the raw number of toggles increases with time spent on trial. On the other hand, a positive significant correlation of `RateTogglingAvailableMsgs` with the answer time in both conditions indicates that the more time is spent on the trial, the more participants tend to toggle.

Last but not least is the `PropTimeOnDist`. It has a significant positive correlation with the answer time only in the Simple trials. This indicates that the more time a participant spends looking at the Distractor, the more time they would spend on the trial in general. This is not surprising as the Distractor is not an important feature in the Simple trials, however, spending more time looking at it would mean that less time is spent looking at the other crucial areas of interest.

The pairwise correlations for the Unambiguous condition were not included in the tables because there were very few significant effects. There were no significant correlations with accuracy, however, the `PropTimeOnComp`, `PropTimeOnDist` and `PropTimeOnAvailableMsgs` have significant positive correlations with the answer time. This is because the Unambiguous trials are very easy to solve, therefore, the participants do not need to look at the Distractor, Competitor or available messages at all. The only feature that has a significant negative correlation with the answer time is `PropTimeOnSentMsg`. This is not surprising as the sent message is perhaps the most important feature in the Unambiguous trials. A more detailed explanation and the table can be found in section A.1.

Conclusion

The pairwise correlations gave us some general understanding about which features are important. The most important finding is, the significant correlation of `PropTimeOnAvailableMsgs` and `RateTogglingAvailableMsgs` to the accuracy in both Simple and Complex conditions. As well as the absence of significant correlation

between the `PropTimeOnDist` and accuracy. This hints us that participants look similarly on the trial in terms of proportional time, but still achieve different accuracy. From here we will move to the linear regression predicting accuracy.

4.3 Predicting Accuracy

The final formula for the model predicting accuracy is presented below. All random effects were removed from the model due to convergence issues:

```
Correct ~ Condition + TrgtPos + Trial +
PropTimeOnTrgt + PropTimeOnComp + PropTimeOnDist +
PropTimeOnSentMsg + PropTimeOnAvailableMsgs +
RateTogglingAvailableMsgs + MsgType + AnswerTime +
Condition:PropTimeOnTrgt + Condition:PropTimeOnComp +
Condition:PropTimeOnDist + Condition:PropTimeOnSentMsg +
Condition:PropTimeOnAvailableMsgs +
Condition:RateTogglingAvailableMsgs +
Condition:AnswerTime +
(1 + Condition + TrgtPos + Trial + PropTimeOnTrgt +
PropTimeOnComp + PropTimeOnDist + PropTimeOnSentMsg +
PropTimeOnAvailableMsgs + RateTogglingAvailableMsgs +
MsgType | Subject)
```

The model had the following encodings for the categorical variables Table 4.3, Table 4.4 and Table 4.5. The Target position can be interpreted as comparing left to center or right and right to center or left for features "TrgtPos2" and "TrgtPos3" respectively. The condition can be interpreted as comparing the Simple condition to the Complex condition and the Unambiguous condition to the Simple and Complex conditions together. The model was trained using the `lme4` package in R. The model was trained using the `glm` function with the following parameters: `family = binomial(link = "logit")`. The resulting coefficients can be seen in Table 4.6.

Message Type	MsgType
Shape	-1
Color	1

Table 4.3: Encoding of the message type categorical variable.

Target Position	TrgtPos2	TrgtPos3
0	1	0
1	0	0
2	0	1

Table 4.4: Encoding of the Target position categorical variable.

Looking at some general findings from Table 4.6, starting from the Intercept, it is positive and significant, indicating that having no information about any of the features, the trial

Condition	Condition1	Condition2
Complex	-1	-1
Simple	1	-1
Unambiguous	0	2

Table 4.5: Encoding of the condition categorical variable.

is more likely to be solved correctly. This is unsurprising as the average accuracy across all trials amounted to 82.5%.

One can see from `Condition1` that the Complex trials are predicted to have significantly lower accuracy comparing to the Simple ones. However, the effect is not fully captured by this coefficient due to the inclusion of the interaction terms. `Condition2` indicates that Unambiguous trials have a higher probability of being solved correctly comparing to Simple and Complex ones, the effect is also significant. So far, the findings fully align with how the trials were designed.

`TrgtPos2` and `TrgtPos3` effects are harder to interpret. The findings suggest that the when the Target is on the left or right side, the probability of correctly answering increases, the effect is significant. This might be somehow related to the fact that the objects are split far apart from each other.

Due to the convergence issues, a corresponding Bayesian model was trained that included all random effects.

The coefficient for `Trial` suggests that people get better as they progress through the experiment. This is expected as the participants have feedback after they answer, which tells whether their answer is correct or not. This finding also aligns with some of the findings from the previous work (Mayn & Demberg, 2023; Mayn et al., 2025).

Moving to the eye-tracking features, the `PropTimeOnComp` has a significant negative effect, indicating that the more participants looks at the Competitor, the more likely they are to answer incorrectly. This effect probably comes from the last look that people do before selecting the object which tips the proportional feature. If a participant cannot derive the correct answer through the reasoning, they will most likely guess among the Target and the Competitor as they both share the sent message feature, making looking at the Competitor negatively correlated to the accuracy.

Furthermore, the `PropTimeOnDist` has a large negative effect, it is almost significant. The effect is most likely coming from the Unambiguous trials which we will discuss later more in details.

The `MsgType` indicates that trials where the sent message is a shape are more likely to be solved correctly. The effect is significant, although relatively small.

Regarding the `AnswerTime`, the general term suggests that the longer one stays on the trial, the more likely they are to answer incorrectly. However, this term should be interpreted together with the interaction terms. We will discuss it further in more details.

The first hypothesis we were interested in was that proportional time on Distractor is positively associated with accuracy on Complex trials, described in subsubsection 2.1.1. Even though the coefficient for the interaction term "Condition1:PropTimeOnDist" is not significant, due to how interaction terms work, we can still interpret how model prediction changes based on the value of the interaction term. First of all, we can take a look at Table 4.7 which shows the trends of the Proportional time on Distractor for each condition. As well as the plot Figure 4.3 which visualizes the trends of the proportional

Predictor	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.23265	0.18460	6.677	2.43e-11 ***
Condition1	0.22087	0.05499	4.017	5.91e-05 ***
Condition2	1.10938	0.14302	7.757	8.69e-15 ***
TrgtPos2	1.92305	0.21183	9.078	< 2e-16 ***
TrgtPos3	1.69844	0.20450	8.305	< 2e-16 ***
Trial	0.22635	0.04825	4.691	2.72e-06 ***
PropTimeOnTrgt	-6.09992	6.15675	-0.991	0.3218
PropTimeOnComp	-12.92581	6.16769	-2.096	0.0361 *
PropTimeOnDist	-11.95582	6.14733	-1.945	0.0518 .
PropTimeOnSentMsg	-9.95567	6.06806	-1.641	0.1009
PropTimeOnAvailableMsgs	-8.69726	6.27784	-1.385	0.1659
MsgType	-0.11506	0.04888	-2.354	0.0186 *
AnswerTime	-0.12537	0.05186	-2.418	0.0156 *
Condition1:PropTimeOnTrgt	-1.93730	1.09696	-1.766	0.0774 .
Condition2:PropTimeOnTrgt	-12.13074	6.12982	-1.979	0.0478 *
Condition1:PropTimeOnComp	-1.30711	1.08177	-1.208	0.2269
Condition2:PropTimeOnComp	-11.30869	6.12510	-1.846	0.0649 .
Condition1:PropTimeOnDist	-1.13168	1.08619	-1.042	0.2975
Condition2:PropTimeOnDist	-12.16270	6.10435	-1.992	0.0463 *
Condition1:PropTimeOnSentMsg	-0.97657	1.09161	-0.895	0.3710
Condition2:PropTimeOnSentMsg	-10.26290	6.03306	-1.701	0.0889 .
Condition1:PropTimeOnAvMsgs	0.70963	1.13710	0.624	0.5326
Condition2:PropTimeOnAvMsgs	-12.08512	6.24131	-1.936	0.0528 .
Condition1:AnswerTime	-0.11566	0.05769	-2.005	0.0450 *
Condition2:AnswerTime	-0.02166	0.03973	-0.545	0.5856

Table 4.6: Summary of the trained model coefficients. Significance levels: . $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Condition	PropTimeOnDist.trend	SE	asymp.LCL	asymp.UCL
Complex	1.339	1.46	-1.53	4.204
Simple	-0.925	1.63	-4.11	2.265
Unambiguous	-36.281	18.30	-72.13	-0.434

Table 4.7: Trends for proportional time on Distractor based on condition.

Contrast	Estimate	SE	z.ratio	p.value
Complex - Simple	2.26	2.17	1.042	0.5504
Complex - Unambiguous	37.62	18.30	2.051	0.1003
Simple - Unambiguous	35.36	18.30	1.927	0.1311

Table 4.8: Contrasts for proportional time on Distractor based on condition.

time on Distractor for each condition. The Table 4.7 indicates the trends we anticipated, however, the columns `asymp.LCL` and `asymp.UCL` indicate the asymptotic lower and upper confidence limits – that is, the 95% confidence interval around the slope estimate. In Simple and Complex cases the interval includes 0, making the slopes not significant. It is worth noting that the slope for the Unambiguous trials is very large and ends up

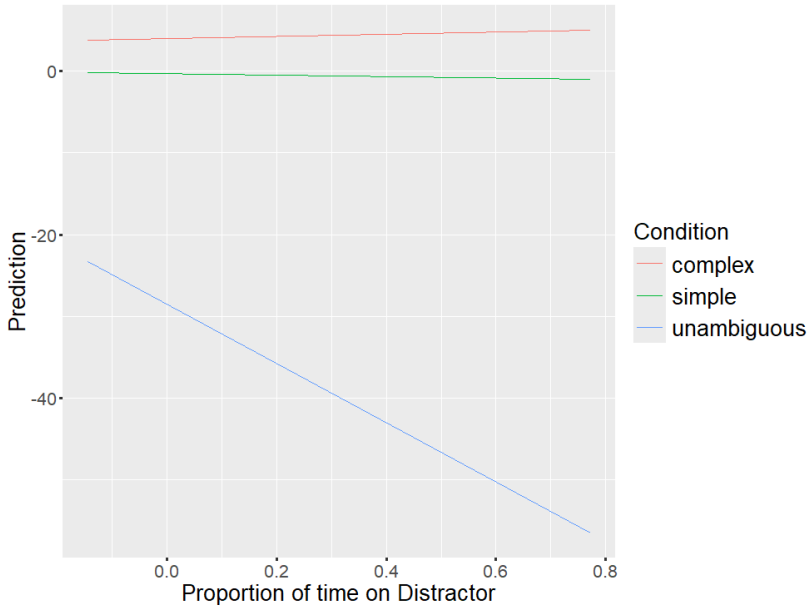


Figure 4.3: Visualization of trends for proportional time on Distractor based on condition.

as significant in the end. However, the main issue is that the amount of incorrectly solved Unambiguous trials is extremely low, as the average accuracy on Unambiguous trials amounted to 98%. This fact makes the slope have very large standard error and subsequently the confidence interval is also very spread. To address the potential that the high uncertainty in the Unambiguous conditions affected model fit and quality, we also fit models to only critical conditions, that is only Simple and Complex ones. Convergence difficulties were found to the same degree, and final model estimates for the trends were comparable, with no changes in significance. We continue with the conclusions we drew from the larger model. Furthermore, we can look at Table 4.8. The table indicates the differences between the slopes and gives corresponding p values for them. The row corresponding to our hypothesis is the first one "Complex - Simple". the findings indicate that the Complex trials will benefit more from the proportional time on Distractor than the Simple ones. However, again, the effect is not significant. In addition, the estimates for the contrasts that include Unambiguous trials are close to being significant, however, these estimates are likely unreliable due to the reasons we discussed earlier.

Condition	**PropTimeOnAvMsgs.trend**	**SE**	**asymp.LCL**	**asymp.UCL**
Complex	2.68	1.51	-0.281	5.64
Simple	4.10	1.70	0.761	7.43
Unambiguous	-32.87	18.70	-69.502	3.77

Table 4.9: Trends for proportional time on available messages based on condition.

The second hypothesis goes as follows, proportional time on available messages is positively associated with accuracy on Simple trials, described in subsubsection 2.1.1. Similarly to how we interpreted the findings for the first hypothesis, we will look at the actual predictions of the model and not solely at the coefficients to capture the whole relation between the interaction terms and the regular ones. The trends are presented

in Table 4.9. From the table we can indeed see a positive effect of proportional time on available messages on Simple condition. Therefore, the findings support the second hypothesis. In addition, the confidence interval does not include 0, suggesting that the effect is significant. As for the Complex and Unambiguous trials, neither of the effects are significant, however, the trend for the Complex trial is also positive, suggesting that the available messages are still important for the participants in the Complex trials. It is worth noting, that the Complex trials can be solved without knowing the available messages at all as was described in subsubsection 2.1.1. Hence, the available messages might not be as important in the Complex condition.

Taking into account the fact that `AnswerTime` and one of the interaction terms including it have significant effects, we will also interpret the trends for this feature. The trends can be seen in Table 4.10. From the table we can conclude that the answer time on the Simple condition has a significant negative effect. This means, that the longer one takes to solve the Simple case, the more likely they are to answer incorrectly. The slopes for the other conditions are not significant.

Condition	**AnswerTime.trend**	**SE**	**asymp.LCL**	**asymp.UCL**
Complex	0.012	0.0678	-0.121	0.1449
Simple	-0.219	0.0938	-0.403	-0.0356
Unambiguous	-0.169	0.1040	-0.372	0.0351

Table 4.10: Trends for Answer Time based on condition.

Predictor	**Estimate**	**Est. Error**	**2.5% CI**	**97.5% CI**
Intercept	0.74	0.25	0.26	1.25
Condition1	0.57	0.12	0.34	0.82
TrgtPos2	2.21	0.33	1.58	2.88
TrgtPos3	1.94	0.32	1.34	2.59
Trial	0.45	0.09	0.28	0.64
PropTimeOnTrgt	6.52	1.41	3.77	9.29
PropTimeOnComp	-2.50	1.29	-5.06	-0.00
PropTimeOnDist	0.21	1.29	-2.33	2.71
PropTimeOnSentMsg	0.84	1.30	-1.71	3.37
PropTimeOnAvailableMsgs	2.74	1.40	-0.01	5.46
RateTogglingAvailableMsgs	1.64	2.35	-2.86	6.29
MsgType1	-0.17	0.09	-0.35	0.00
AnswerTime	-0.20	0.09	-0.37	-0.03
Condition1:PropTimeOnTrgt	-2.59	1.27	-5.08	-0.12
Condition1:PropTimeOnComp	-1.81	1.26	-4.31	0.64
Condition1:PropTimeOnDist	-1.43	1.26	-3.91	1.03
Condition1:PropTimeOnSentMsg	-1.08	1.27	-3.61	1.36
Condition1:PropTimeOnAvailableMsgs	-0.52	1.37	-3.25	2.15
Condition1:RateTogglingAvailableMsgs	4.24	2.29	-0.21	8.70
Condition1:AnswerTime	-0.12	0.08	-0.28	0.04

Table 4.11: Bayesian logistic regression coefficients. All $\hat{R}$ values were equal to 1, indicating convergence.

The decomposition by condition was observed similarly to the MLE model. However, in the case of Bayesian model, none of the interesting effects were found to be significant.

Conclusion

The first hypothesis with the Distractor was not confirmed by the main model. That is, the proportional time on Distractor is not positively associated with the accuracy on Complex trials. On the other hand, the proportional time on available messages is significantly positively associated with the accuracy on Simple trials.

The full conclusion will be drawn later, however, the general pattern suggests that participants fail to solve the trial not due to the lack of information, but rather due to lack of depth during the reasoning process. While the proportional time on available messages might to some extent describe the reasoning process.

4.4 On Distractor

The model presented here predicts the likelihood (binary outcome) of participants gazing at the Distractor during correctly solved trials. The final formula for the model predicting the likelihood of looking at the Distractor for the correctly solved trials is presented below. Most of the random effects were removed from the model due to convergence issues:

```
OnDist ~ Condition + Trial + MsgType + TrgtPos +
    (1 | Subject)
```

Coefficient	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.056241	0.039379	-77.611	<2e-16 ***
Condition1	0.013373	0.006440	2.076	0.0379 *
Condition2	-0.086425	0.004691	-18.425	<2e-16 ***
Trial	-0.053649	0.005713	-9.391	<2e-16 ***
MsgType	-0.017492	0.005713	-3.062	0.0022 **
TrgtPos2	1.772965	0.018094	97.986	<2e-16 ***
TrgtPos3	1.795716	0.018151	98.931	<2e-16 ***

Table 4.12: Summary of the model predicting the likelihood of looking at the Distractor for the correctly solved trials. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In this case, no interaction terms were included in the model. Hence, the model coefficients can be interpreted directly as each coefficient is the effect of the corresponding feature that appears only once in the whole model. The encodings for the categorical variables are the same as in Table 4.3, Table 4.4 and Table 4.5. The model was trained using the `lme4` package in R. The model was trained using the `glmer` function with the following parameters: `family = binomial(link = "logit")`. The resulting coefficients can be seen in Table 4.12.

Starting from the Intercept, it is negative and significant, indicating that without any information about the correctly solved trial, the predicted gaze point is likely to be not on the Distractor. This is unsurprising as the average time spent looking at the Distractor was 14.3% across all correctly solved trials.

`Condition1` indicates that the Distractor is more likely to be looked at in the Simple trials than in the Complex ones. This finding is the opposite of what we expected according to the hypothesis 3 described in subsubsection 2.1.2. The effect is small but

significant. This indicates that taking into account the correctly solved trials, not only the Distractor was not looked at more in the Complex trials, but it was actually looked at more in the Simple trials. A possible explanation is that a skilled participant would apply the same strategy to both the Simple and Complex trials. `Condition2` indicates that the Distractor is less likely to be looked at in the Unambiguous trials than in the Simple and Complex ones. The effect is significant and large, indicating that the Distractor is not looked at as much in the Unambiguous trials. This finding is not surprising as the Distractor is not important in the Unambiguous trials. More importantly this implies that skilled participants tend to use different strategies in Unambiguous trials comparing to Simple and Complex ones.

`Trial` indicates that the Distractor is less likely to be looked at as the trial number increases. This is expected as the participants get more efficient as they progress through the experiment. While closer to the beginning they would explore the trials more thoroughly, later on they would be more likely to look at other areas of interest and not at the Distractor, which is not as important.

`MsgType` indicates that the Distractor is less likely to be looked at when the sent message is a color. The effect is significant but relatively small. We were unable to find a reasonable explanation for this effect.

`TrgtPos2` and `TrgtPos3` indicate that the Distractor is more likely to be looked at when the Target is on the left or right side. The effect is significant and large. There are two possible explanations for that. The first one is that the participants are more likely to look at the center by default and the Distractor is more likely to be located in the center when the Target is on the left or right side. The second one is that due to the fact that the central area is the closest to other areas of interest and, therefore, is more likely to get a gaze point there due to inaccuracy of WebGazer. In this case, the gaze points could be coming from the available messages, sent message or any of the other main objects on the screen. By the same logic, if the Target is on the left or right, the Distractor is more likely to be in the center and, therefore, more likely to get a gaze point.

Conclusion

The hypothesis 3 did not hold according to the model. The fact that the Distractor was not looked at more in the Complex trials can be explained by the fact that a skilled participant would apply the same strategy to both the Simple and Complex trials. Even though, the Distractor is important in the Complex trials, as was discussed in section 1.2.

4.5 On Available Messages

The model presented here predicts the likelihood (binary outcome) of participants gazing at the bank of available messages during correctly solved trials. The final formula for the model predicting the likelihood of looking at the available messages for the correctly solved trials is presented below. Most of the random effects were removed from the model due to convergence issues:

```
OnAvMsgs ~ Condition + Trial + MsgType + TrgtPos +
    (1 + TrgtPos | Subject)
```

The model was trained using the `lme4` package in R. The model was trained using

Coefficient	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.330044	0.071251	-32.702	<2e-16 ***
Condition1	-0.063455	0.007056	-8.993	<2e-16 ***
Condition2	-0.385246	0.006972	-55.260	<2e-16 ***
Trial	-0.085418	0.006603	-12.936	<2e-16 ***
MsgType	0.085259	0.006573	12.972	<2e-16 ***
TrgtPos2	-0.143025	0.072737	-1.966	0.0493 *
TrgtPos3	-0.140503	0.088966	-1.579	0.1143

Table 4.13: Summary of the model predicting the likelihood of looking at the available messages for the correctly solved trials.Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

the `glmer` function with the following parameters: `family = binomial(link = "logit")`. The resulting coefficients can be seen in Table 4.13. The encodings for the categorical variables are the same as in Table 4.3, Table 4.4 and Table 4.5.

The Intercept is negative and significant, indicating that without any information about the correctly solved trial, the predicted gaze point is likely to be not on the available messages. This is unsurprising as the average time spent looking at the available messages was 9% across all correctly solved trials.

`Condition1` indicates that the available messages are more likely to be looked at in the Complex trials than in the Simple ones. The effect is small but significant. This does not align with our hypothesis 4 described in subsubsection 2.1.2. This can be explained similarly to why the the hypothesis 3 did not hold. Skilled participants tend to use a similar approach to both the Simple and Complex trials. However, clearly, the available messages are still important for the participants in the Complex condition even though they could theoretically be solved without the available messages. `Condition2` indicates that the available messages are less likely to be looked at in the Unambiguous trials than in the Simple and Complex ones. The effect is significant and large, indicating that the available messages are not looked at as much in the Unambiguous trials. This finding is expected as the available messages are not important in the Unambiguous trials.

Similarly to the Distractor model, `Trial` indicates that the available messages are less likely to be looked at as the trial number increases. This is expected as the participants get more efficient as they progress through the experiment. While closer to the beginning they would explore the trials more thoroughly, later on they would be adapt a more efficient strategy, which would decrease the time spent looking at the available messages.

`MsgType` indicates that the available messages are less likely to be looked at when the sent message is a shape. The effect is significant. A possible explanation is that when the message is a shape, the Competitor would differ from the Target by color. Noticing the distinct feature of the Competitor is crucial when solving the Simple trials. The color might be easier to see with a peripheral vision than a shape. This would make the available messages more likely to be looked at when the sent message is a shape.

As for the Target position, this time it does not play as important role as in the Distractor model. However, taking into account the findings from the Distractor model, it would be reasonable to assume that the Target position in the middle would make the available messages more likely to get a gaze point. However, the effect is not significant for the `TrgtPos3` and is relatively low for both features.

Conclusion

While the hypothesis 4 did not hold according to the model, we still got important pieces of information. Mainly, similarly to the previous model, that predicted fixations on Distractor, we saw that the skilled participants do not look at the available messages more during the Simple trials comparing to Complex ones. The following is rather the case, the skilled participants tend to keep a similar attention profile in both Simple and Complex conditions, but adapt a different strategy for the Unambiguous trials.

4.6 Rate of Toggling

The model presented here predicts the rate of toggling between the available messages during correctly solved trials. The final formula for the model predicting the rate of toggling is presented below.

```
RateTogglingAvailableMsgs ~ Condition + TrgtPos + Trial +
StrategyLabel + Correct + MsgType + AnswerTime +
Condition:StrategyLabel + Condition:AnswerTime +
(1 + Condition + TrgtPos + Trial +
StrategyLabel + Correct + MsgType + AnswerTime | Subject)
```

where `StrategyLabel` is a categorical variable that indicates the strategy label of the participant, which was derived from the annotations of the strategies that participants entered in the end of the experiment. The `Correct` variable indicates whether the trial was solved correctly or not. The model was trained using the `brms` package in R. The model was trained using the `brm` function. The resulting coefficients can be seen in Table 4.14.

Predictor	Estimate	Est. Error	2.5% CI	95% CI
Intercept	0.00161	0.00238	-0.00304	0.00629
Condition1	0.00035	0.00078	-0.00118	0.00186
TrgtPos2	-0.00122	0.00169	-0.00451	0.00206
TrgtPos3	-0.00058	0.00173	-0.00398	0.00276
Trial	0.00152	0.00088	-0.00021	0.00327
StrategyLabel1	0.00560	0.00230	0.00102	0.01009
StrategyLabel2	0.00317	0.00105	0.00110	0.00523
Correct	0.00062	0.00184	-0.00300	0.00426
MsgType1	0.00136	0.00071	-0.00003	0.00276
AnswerTime	0.01541	0.00197	0.01171	0.01948
Condition1:StrategyLabel1	0.00182	0.00106	-0.00028	0.00392
Condition1:StrategyLabel2	0.00092	0.00047	-0.00002	0.00184
Condition1:AnswerTime	0.00057	0.00084	-0.00108	0.00223

Table 4.14: Regression coefficients from the model. All $\hat{R}$ values were close to 1, indicating convergence.

The coefficients demonstrate that the rate of toggling is positively associated with the `AnswerTime` feature, which is also supported by the correlation analysis in subsection 3.2.4. Moreover, the general coefficients of `StrategyLabel` are significant and

indicate that more skilled participants according to the labels toggle with a higher rate than less skilled ones. Furthermore, one can look at the decomposition according to condition in Table 4.15. The results indicate that the rate of toggling has a significant negative effect for L_0 listeners in Simple condition. As well as positive significant effect for L_2 listeners in both Simple and Complex conditions. Further supporting the idea that more skilled participants toggle more often than less skilled ones.

Condition	StrategyLabel	emmean	lower.HPD	upper.HPD
complex	0	-0.00505	-0.01208	0.00175
simple	0	-0.00983	-0.01690	-0.00315
complex	1	0.00253	-0.00458	0.00975
simple	1	0.00506	-0.00228	0.01228
complex	2	0.00547	0.00089	0.01023
simple	2	0.00987	0.00495	0.01443

Table 4.15: Estimated marginal means (emmean) with 95% HPD intervals by condition and strategy label.

4.7 Strategies

As was discussed in subsection 3.1.1 participants completed 2 strategy trials in the end of the experiment. One of the Simple condition and one of the Complex one. We have annotated them similarly to how it was done in Mayn and Demberg (2023). Based on the annotations the labels corresponding to L_0, L_1 or L_2 listeners were assigned. The labels were assigned as follows, if the Complex trial is annotated as "correct_reasoning", participant gets an L_2 label. If only the Simple trial is annotated with "correct_reasoning", the participant would get an L_1 label. If neither of the previous were assigned, the participant gets the label L_0. Based on the labeling, the following plots were made Figure 4.4, Figure 4.5 and Figure 4.6. In order to match the scale of other features, the `MeanAnswerTime` feature was divided by 10 000, hence, value of one corresponds to spending 10 seconds on the trial.

While the labeling was done solely based on the annotations of strategies that participants entered, there is a clear alignment between the listeners from RSA and the derived labels. That is, L_2 listeners are capable of solving trials from all 3 conditions, while L_1 fail only the trials from the Complex condition and L_0 are only capable of solving the trials from Unambiguous condition.

Furthermore, the plots show similar patterns to the models. For example, there is an increase in `PropTimeOnAvailableMsgs` as the label goes from L_0 to L_1 and to L_2. Indicating that more skilled participants comparing to less skilled ones, spend more time looking at the available messages, similarly to the findings from the main model in section 4.3. Again, we do not see any differences for the `PropTimeOnDist`, also aligning with the main model. In addition, the Unambiguous plot in Figure 4.6 suggests that participants do not vary in their attention profile in Unambiguous condition.

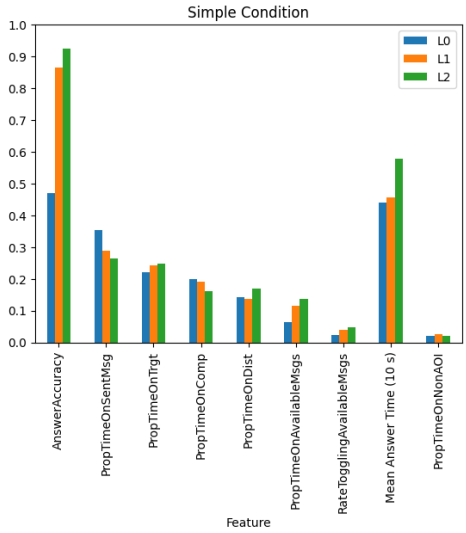


Figure 4.4: Comparison of features for different labels on trials of Simple condition.

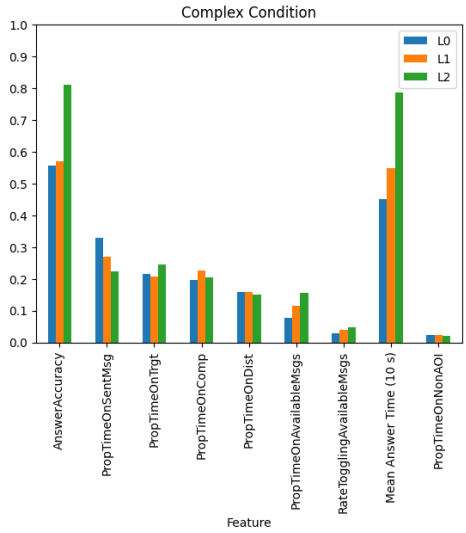


Figure 4.5: Comparison of features for different labels on trials of Complex condition.

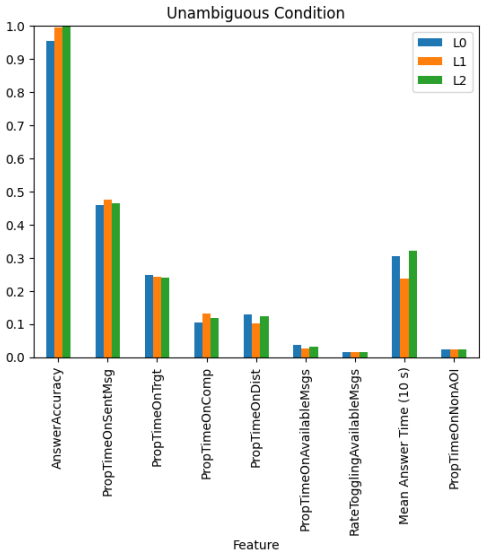


Figure 4.6: Comparison of features for different labels on trials of Unambiguous condition.

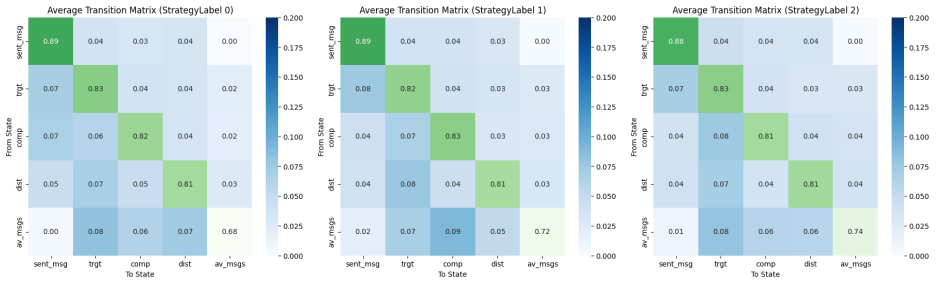


Figure 4.7: Average transition matrices for the L_0, L_1 and L_2 listeners.

4.8 Scanpath Analysis

The scanpath analysis is a difficult analysis as the data itself has different lengths. This means that one cannot directly use similar models as we did for the other features. Hence, we resort to the following approach of modeling the scanpaths as Markov Models.

4.8.1 Markov Model

The Markov Model is a statistical model that describes the scanpath with a matrix of transition conditional probabilities. That is, the model describes the probability of transitioning from one area of interest to another. The approach is similar to the one used by Coutrot, Hsiao, and Chan (2018). The main difference is that we do not use the hidden Markov Model, as we defined the areas of interest ourselves. Figure 4.7 presents the transition matrices for the L_0, L_1 and L_2 listeners. It is important to note that due to the fact that no fixation detection was done, the diagonal elements of the matrices are very high comparing to other elements. This is expected as the participants are likely to look at the same area of interest multiple times during one fixation which is not compressed to one event.

The resulting entries of transition matrices were used as features to predict the labels of the listeners. Due to the three classes of listeners, we used tree classifiers, one for each critical condition. The trees' hyperparameters were tuned using the `GridSearchCV` function from the `sklearn` package in Python. In order to keep interpretability of the results, we limited the search space of max_depth up to 4. The cross validation had 5 folds. The best mean f1 weighted score for the Simple condition amounted at 0.47 and for the Complex condition at 0.51. Further we present the visualization of the trees trained on the whole data using the best found hyperparameters in Figure 4.8. The trees were trained based on Gini impurity. It can be seen at every node indicating its purity. The top part of the node indicates the feature that was used to split the data at that node. Furthermore, the value gives us insight into the current distribution of the labels at that node. The higher the value for a certain class is the more likely the class is to be predicted at that node.

Even though the scores are not particularly impressive, we would like to interpret some of the splits made by the decision trees along the way. The Simple condition tree in Figure 4.8 indicates that the the available messages features allow for good splits in the data, firstly identifying L_0 listeners with low `comp_to_av_msgs` values. And yet another split on the same feature allows to identify L_2 listeners with high `comp_to_av_msgs` values.

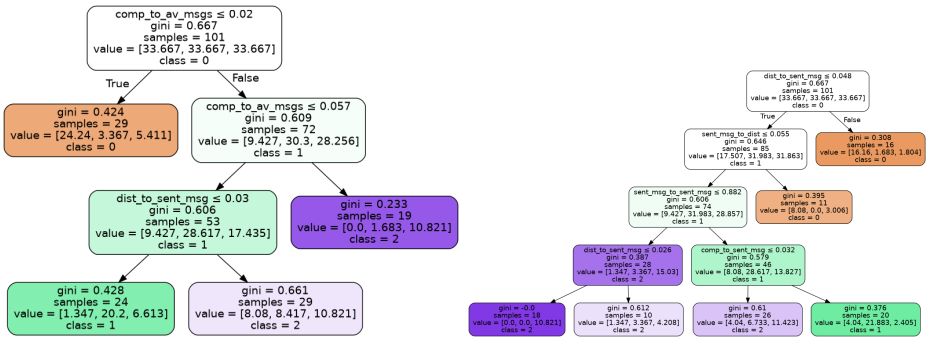


Figure 4.8: Scanpath based trees for the Simple (left) and Complex (right) conditions.

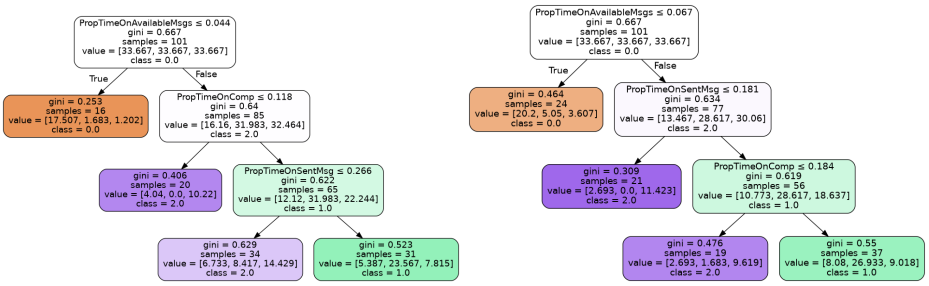


Figure 4.9: `PropTimeOn` based trees for the Simple (left) and Complex (right) conditions based on proportional features.

The Complex condition tree, on the other hand, involves Distractor features at the top levels. The root node splits on `dist_to_sent_msg` feature, while the following node splits on `sent_msg_to_dist`, low values at both splits indicate L_0 listeners. It is not particularly clear how the Distractor and sent message interact in this setting, however, the Distractor is clearly important for the Complex condition.

Furthermore, we compared the features derived from transition matrices to the proportional features we used in the main model. The best mean f1 weighted score for the Simple condition amounted at 0.5 and for the Complex condition at 0.56. Potentially indicating that the proportional features are more informative than the transition matrices. The visualizations of the trees trained on the proportional features can be seen in Figure 4.9.

The trees in this case look very similar for both conditions, potentially indicating that the transition matrices were able to make a better distinguish between them. Available messages are again found to be important, in this case, for both conditions, being on the very top of the tree, their split distinguishes L_0 listeners from L_1 and L_2. In contrast to the scanpath based trees, the Distractor is not found to be important in this case.

4.9 ACT-R Model

Chapter 5
Conclusion and General Discussion

5.1 Conclusion

In this thesis, we approached the research question from multiple analytical angles to better understand how visual attention relates to reasoning in reference games. First, we conducted a pairwise correlation analysis between eye-tracking features and two key trial-level outcomes: accuracy and response time. Next, we built a logistic regression model to predict whether a given trial would be solved correctly, based on the participant's gaze behavior. To further investigate the role of attention, we implemented two additional logistic regression models that focused on specific fixation patterns: one predicted whether a gaze point would land on the Distractor, and the other did the same for the bank of available messages. Finally, we analyzed participants' self-reported strategies and labeled them according to the Rational Speech Act (RSA) framework (L_0, L_1, L_2). This allowed us to compare how different reasoning strategies manifest in distinct attentional profiles.

Although participants seem to grasp the problem in a similar way, their answers still differ quite a lot. We saw this from the fact that the Distractor did not turn out to be associated with accuracy, even though it is a very important piece of information for the Complex trials. Thus, participants fail the trials not because they miss an important piece of information as we hypothesized in section 1.2, where we showed that not looking at the Distractor would make the Complex condition unsolvable. The problem appears to lie in how participants interpret or apply the information they attend to, rather than in what information they access.

The consistent association between accuracy and proportional time spent on the bank of available messages suggests that this behavior may reflect more than simple visual scanning. It is plausible that participants engage in matching messages to objects, or in other forms of referential reasoning. In the context of reference games, such (possibly recursive) reasoning processes are essential. Thus, increased attention to the available messages may serve as an indicator of deeper or more robust reasoning, ultimately leading to higher accuracy.

While we could not directly infer participants' reasoning strategies, we were able to identify eye-tracking features associated with trial accuracy. Further research might include more sophisticated analysis to identify the actual strategies of the participants. The main finding suggests that the bank of available messages is an important area of interest when predicting whether a person is going to solve the trial correctly. A possible future work might include detecting which messages the participants are looking at and how they are using them. This reflects a design limitation: anticipating potential quality issues with WebGazer, we placed the available messages close together – thus sacrificing the ability to detect which message the participants are looking at. However, this is a very interesting question and could be a good future work.

Ongoing research on modeling participant behavior in reference games using ACT-R (Duff, Mayn, & Demberg, 2025) may provide deeper insights into the cognitive strategies underlying their responses.

This thesis contributes to a wider ongoing research on reference games. Overall, this work demonstrates that even with limited-resolution eye-tracking data, meaningful patterns can be extracted that shed light on participants' reasoning processes in reference games. This thesis provides a strong foundation for future work that would include eye-tracking data in reference games or another type of experiment with a similar setup.

5.2 General Discussion

5.2.1 Scanpaths

In this study, did not account for the order of the gaze points. However, the scanpath by itself includes this information. We were unable to come up with a good strategy how one could encode the scanpath in a way that would be useful and feasible for the analysis. The scanpaths are not fixed length, posing a problem that every entry would be of a different length which would make it impossible to use them in a model. One idea could be to use RNNs to analyze the scanpaths, however, this would require a lot of data and a lot of tuning. We did a few attempts in training CNNs on the generated plots of the scanpaths, however, the results were not promising and due to low interpretability of the model and time constraints we decided to focus on the more traditional features.

5.2.2 Peripheral Vision

It is possible that the participants used their peripheral vision to look at the objects. If one specifically tries to solve the trial without moving the eyes, one could still do it. This definitely poses a problem for the eye-tracking data, as it would be impossible to detect this kind of behavior. We do not have any evidence that this is the case, but it is a possibility. People would not specifically try to solve the trials without the eye movements, however, it is quite probable that they capture some features such as color of the Distractor in case that the sent message is a shape during the trial. This is somewhat supported by the likelihood model predicting whether a gaze point would end up on the Distractor or not section 4.4. `MsgType` had a very low but significant coefficient indicating exactly this behavior. Although this is not a strong evidence for the peripheral vision, the potential problem should be taken into account.

5.2.3 Toggling

We only implemented features that describe proportional time spent on the area of interest. However, as we described in the section 5.1, an increase in proportional time on available messages might indicate a deeper reasoning process. One could also think about defining a feature to describe the toggling between the available messages and other areas of interest. Similarly to how it was done in Vigneau et al. (2006). The toggling might be an important predictor for accuracy if one makes repeated matches of the messages and the objects.

5.2.4 Proportional Eye Tracking Features

We suspect that the proportional eye-tracking features might be influenced by the very last look at one of the objects. Therefore, making them good predictors of accuracy, as we saw in section 4.3 or section 4.2. The Competitor and Distractor turned out to be decisive for the accuracy of the predictions. This indicates that people look more at the object they would choose. As for the future work, one should consider trimming the last saccade before the decision, this should make the proportional features more reliable. On the other hand, if this does not solve the issue, this would indicate that people tend to make a decision not in the end of the trial. This would also be an interesting finding.

The way we assigned the fixations to the stimuli is not optimal as we clearly disregarded the timings of the gaze predictions as well as added some of the gaze points that are from saccades rather than from fixations. Both of the problems could have been solved with a fixation detection algorithm. A more advanced fixation detection algorithm could be implemented in the future to improve the preprocessing part of the data analysis. Currently as was discussed in subsection 3.2.3 the fixation detection algorithm falls short in the case of low sampling rates that WebGazer demonstrates. The algorithm is unable to detect some fixations in the data, which can lead to a loss of information.

5.2.5 Eye Tracking

The eye-tracking data collected by WebGazer was definitely worse than one could collect in the lab. However, clearly this thesis is an example that it is possible to conduct an eye-tracking experiment with WebGazer. The data is not perfect, but it is good enough to be able to draw some conclusions from it. However, the library is far from perfect and definitely needs a lot of adjustments to the specific setups one could be interested in. Designing the experiment took far more time than the actual analysis of the data. Each participant has a different setup. And, although, the calibration of the eye tracker must be conducted quite strictly, it is also important to make sure that the calibration is doable for the participants. We have achieved some balance between the two, but there is still a lot of room for improvement.

On the other hand, WebGazer has a few clear advantages. First of all, the fact that it is possible to collect data from a large number of participants. This is something that is not possible in the lab, especially for an eye-tracking study as it would require individual treatment of each participant, as oppose to each participant working in parallel in the online experiment. Secondly, WebGazer is free and open source, which makes it available for everyone. This is a clear advantage over the lab eye trackers that are expensive and not available for everyone.

Appendices

Appendix A
Appendix A: Additional Details

A.1 Pairwise Correlations

Feature	Mean	SD	Accuracy	Mean Answer Time
PropTimeOnSentMsg	0.46	0.11	0.12	-0.38 ***
PropTimeOnAvailableMsgs	0.03	0.03	-0.1	0.43 ***
PropTimeOnTrgt	0.24	0.08	-0.04	-0.0
PropTimeOnDist	0.12	0.06	-0.13	0.25 *
PropTimeOnComp	0.12	0.06	-0.01	0.2 *
PropTimeOnNonAOI	0.02	0.02	0.11	0.15
MeanAnswerTime	3005.96	2179.31	-0.1	—
AnswerAccuracy	0.99	0.05	—	—

Table A.1: Unambiguous condition. Correlation table showing the relationships between features, accuracy, and mean answer time. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The table above shows the pairwise correlation between the features, accuracy, and mean answer time in the Unambiguous trials. There were no significant correlations between the features and accuracy. The only significant correlations were in the Mean Answer Time column, the results are not surprising as the Unambiguous trials are not difficult to solve, and most of the participants adopted a similar strategy where they would just match the sent message with the only object that has the same feature. This strategy involves only looking at sent message and the Target, making any deviations indicate that the participant is not following the strategy and they will take longer on the trial. It is worth to mention, that while the strategy is very efficient, a deviation from it does not make one answer incorrectly as can be seen from the results.

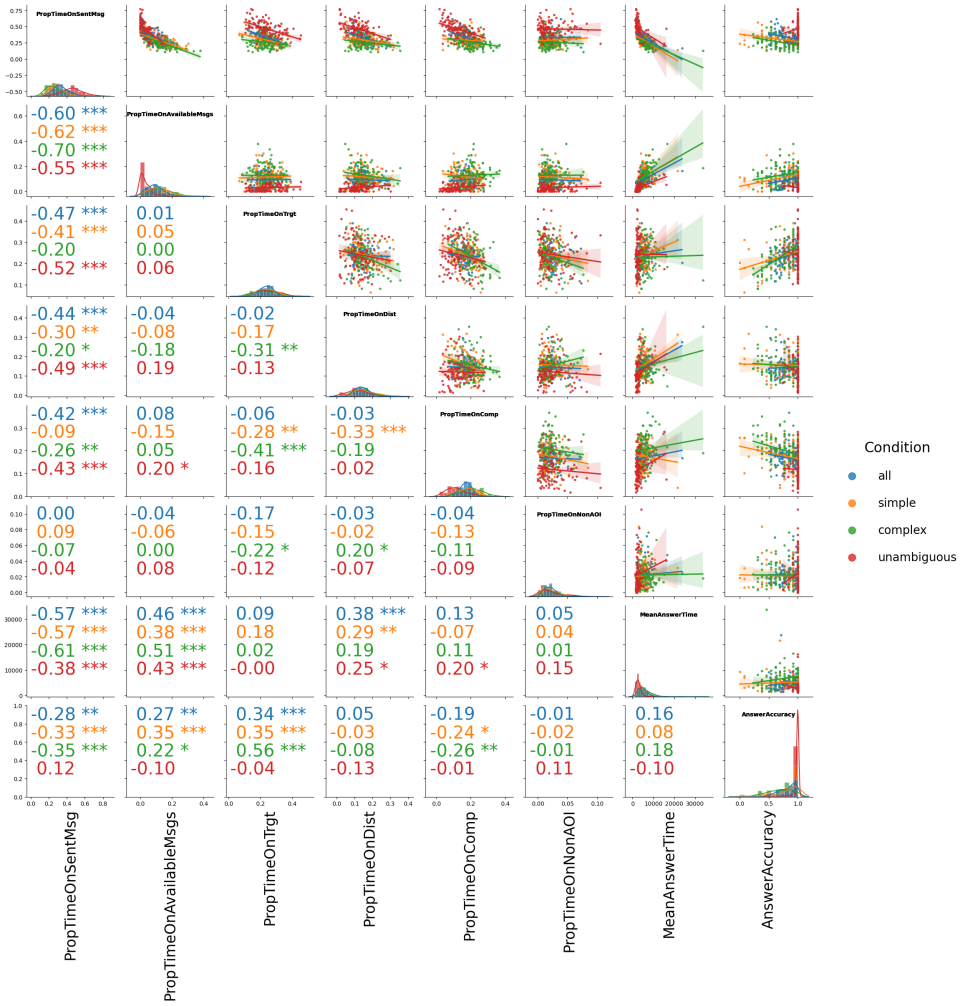


Figure A.1: Full pairwise correlation matrix for all conditions and features. The diagonal shows the distribution of the features. The top right triangle shows the scatter plots along with the regression lines. The bottom left triangle shows the correlation coefficients. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix B

Appendix B: Supplementary Material

B.1 Trials

ID	Condition	Sent Msg	Trgt	Comp	Dist	Msg1	Msg2	Msg3	Msg4
1	simple	ci	ci_bl	ci_gr	tr_re	ci	tr	re	gr
2	simple	re	tr_re	sq_re	ci_bl	re	bl	ci	sq
3	simple	ci	ci_bl	ci_gr	sq_re	ci	sq	re	gr
4	simple	re	sq_re	tr_re	ci_gr	re	gr	ci	tr
5	simple	sq	sq_bl	sq_gr	tr_re	sq	tr	re	gr
6	simple	gr	tr_gr	ci_gr	sq_bl	gr	bl	sq	ci
7	simple	sq	sq_bl	sq_gr	ci_re	sq	ci	re	gr
8	simple	gr	ci_gr	tr_gr	sq_re	gr	re	sq	tr
9	simple	tr	tr_bl	tr_gr	sq_re	tr	sq	re	gr
10	simple	bl	sq_bl	ci_bl	tr_gr	bl	gr	tr	ci
11	simple	tr	tr_bl	tr_gr	ci_re	tr	ci	re	gr
12	simple	bl	ci_bl	sq_bl	tr_re	bl	re	tr	sq
13	complex	ci	ci_gr	ci_re	sq_gr	ci	tr	re	gr
14	complex	re	sq_re	ci_re	sq_gr	re	bl	ci	sq
15	complex	ci	ci_gr	ci_re	tr_gr	ci	sq	re	gr
16	complex	re	tr_re	ci_re	tr_bl	re	gr	ci	tr
17	complex	sq	sq_gr	sq_re	ci_gr	sq	tr	re	gr
18	complex	gr	ci_gr	sq_gr	ci_re	gr	bl	sq	ci
19	complex	sq	sq_gr	sq_re	tr_gr	sq	ci	re	gr
20	complex	gr	tr_gr	sq_gr	tr_bl	gr	re	sq	tr
21	complex	tr	tr_gr	tr_re	ci_gr	tr	sq	re	gr
22	complex	bl	ci_bl	tr_bl	ci_re	bl	gr	tr	ci
23	complex	tr	tr_gr	tr_re	sq_gr	tr	ci	re	gr
24	complex	bl	sq_bl	tr_bl	sq_gr	bl	re	tr	sq
25	unambiguous	ci	ci_re	sq_re	sq_bl	ci	sq	re	gr
26	unambiguous	re	ci_re	ci_gr	tr_gr	re	gr	ci	sq
27	unambiguous	ci	ci_re	tr_re	tr_bl	ci	tr	re	gr
28	unambiguous	re	ci_re	ci_bl	sq_bl	re	bl	ci	tr
29	unambiguous	sq	sq_re	ci_re	ci_bl	sq	ci	re	gr
30	unambiguous	gr	sq_gr	sq_re	tr_re	gr	re	sq	ci
31	unambiguous	sq	sq_re	tr_re	tr_bl	sq	tr	re	gr
32	unambiguous	gr	sq_gr	sq_bl	ci_bl	gr	bl	sq	tr
33	unambiguous	tr	tr_re	ci_re	ci_bl	tr	ci	re	gr
34	unambiguous	bl	tr_bl	tr_re	sq_re	bl	re	tr	ci
35	unambiguous	tr	tr_re	sq_re	sq_bl	tr	sq	re	gr
36	unambiguous	bl	tr_bl	tr_gr	ci_gr	bl	gr	tr	sq
37	strtgy_simple	ci	ci_bl	ci_gr	tr_re	ci	tr	re	gr
38	strtgy_complex	bl	sq_bl	tr_bl	sq_gr	bl	re	tr	sq

Table B.1: Table of all trials used in the experiment.

References

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205-238. Retrieved from https://www.sciencedirect.com/science/article/pii/0160289684900096 doi: https://doi.org/10.1016/0160-2896(84)90009-6

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven progressive matrices test. *Psychological Review*, *97*, 404-431. doi: 10.1037/0033-295x.97.3.404

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with hidden markov models. *Behavior Research Methods*, *50*(1), 362–379. Retrieved from https://doi.org/10.3758/s13428-017-0876-8 doi: 10.3758/s13428-017-0876-8

Duff, J., Mayn, A., & Demberg, V. (2025). An act-r model of resource-rational performance in a pragmatic signaling game.

Frank, M. C., Emilsson, A. G., Peloquin, B., Goodman, N. D., & Potts, C. (2016). *Rational speech act models of pragmatic reasoning in reference games.* PsyArXiv. Retrieved from osf.io/preprints/psyarxiv/f9y6b doi: 10.31234/osf.io/f9y6b

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998-998. Retrieved from https://www.science.org/doi/abs/10.1126/science.1218633 doi: 10.1126/science.1218633

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual- vs. population-level probabilistic modeling. *PLOS ONE*, *11*(5), 1-25. Retrieved from https://doi.org/10.1371/journal.pone.0154854 doi: 10.1371/journal.pone.0154854

Grice, H. P. (1975). Logic and conversation. *Syntax and semantics*, *3*, 43–58.

Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, MA, USA: Wiley-Blackwell.

Mayn, A., & Demberg, V. (2023, 06). High performance on a pragmatic task may not be the result of successful reasoning: On the importance of eliciting participants' reasoning strategies. *Open Mind*, *7*, 156-178. Retrieved from https://doi.org/10.1162/opmi_a_00077 doi: 10.1162/opmi_a_00077

Mayn, A., Loy, J. E., & Demberg, V. (2025, 01). Beliefs about the speaker's reasoning ability influence pragmatic interpretation: Children and adults as speakers. *Open Mind*, *9*, 89-120. Retrieved from https://doi.org/10.1162/opmi_a_00180 doi: 10.1162/opmi_a_00180

Mykhalievskyi, T. (2025). *Source code for thesis project.* Retrieved from https://github.com/GAtemROU/Thesis-Project

Mykhalievskyi, T., Duff, J., & Demberg, V. (2025). *Preregistration for thesis project.* Retrieved from https://osf.io/hf4vs/?view_only=0d1880024b7349c4aec25b5bc318c85e

Palan, S., & Schitter, C. (2018). Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22-27. Retrieved from https://www.sciencedirect.com/science/article/pii/S2214635017300989 doi: https://doi.org/10.1016/j.jbef.2017.12.004

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the 25th international joint conference on artificial intelligence (ijcai)* (pp. 3839–3845).

Pusse, F., Sayeed, A., & Demberg, V. (2016). Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations* (pp. 57–61).

Reback, J., McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Klein, A., ... others (2020). pandas-dev/pandas: Pandas 1.0. 5. *Zenodo*.

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. Retrieved from https://doi.org/10.3758/s13428-017-0913-7 doi: 10.3758/s13428-017-0913-7

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*(3), 261-272. Retrieved from https://www.sciencedirect.com/science/article/pii/S0160289605001248 doi: https://doi.org/10.1016/j.intell.2005.11.003

von der Malsburg, T. (2015, oct). *Saccades: An r package for detecting fixations in raw eye tracking data.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.31799 doi: 10.5281/zenodo.31799

Wisiecka, K., Krejtz, K., Krejtz, I., Sromek, D., Cellary, A., Lewandowska, B., & Duchowski, A. (2022). Comparison of webcam and remote eye tracking. In *2022 symposium on eye tracking research and applications.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3517031.3529615 doi: 10.1145/3517031.3529615