Investigating Reasoning with Hypotheses: The RIP2 Corpus

Ella Schad, Clara Seyfried, Chris Reed
University of Dundee, Dundee DD1 4HN, UK
{e.m.schad, c.seyfried, c.a.reed}@dundee.ac.uk

Abstract

Analyses of hypothesis generation in fictionalised environments have significant potential for exploring factors influencing reasoning and decision-making in naturalistic contexts. Based on transcripts of 16 groups playing a murder mystery game, with a total of 42 human participants, RIP2 is a 177,000 word corpus exemplifying reasoning in the forensic domain. With a 80,000 word representative sample of the corpus annotated using an argumentation framework, RIP2 is nearly twice the size of the RIP Corpus of Collaborative Hypothesis-Making (RIP1), currently the only existing corpus of hypothesis-making in group environments. With an new experimental set-up and guidelines for annotating both cases of hypothesising and conjecturing, RIP2 offers insight into how participants generate, maintain, and reject hypotheses, as well as how they interact with others' contributions. Based on its close exploration of six groups (three successful), this corpus particularly allows for group-level comparisons of factors influencing group success. Within this paper, we discuss the main contributions for understanding hypothesising and collaborative reasoning, and offer use cases for extended work demonstrating how analysis of hypothesis generation can be used for future research on argumentation quality and decision-making.

Keywords: argumentation, collaborative reasoning, corpus, game, hypotheses, problem solving

1. Introduction

RIP2 (the second Resolving Investigative hyPotheses corpus) is a 177,000 word study into how humans develop and reason with hypotheses in the context of an experimental setting. Groups of up to three participants met for the first time to play a murder mystery game, offering rich transcripts of natural language data, here annotated using the annotation framework Inference Anchoring Theory (Budzynska et al., 2014a). By empirically investigating argumentation in a fictionalised forensic environment, with an annotated subset of the corpus of 80,000 words, which includes 2666 relations, we are able to show the myriad ways in which humans make, maintain, and interact with hypotheses. Additionally, we can link this work with group dynamic research, showing in which contexts groups best succeed in their joint reasoning.

We build on the work of Schad et al. (2024b) and develop the largest annotated hypothesising corpus to date, linked to a wider resource which includes additional transcripts and is supplemented with background measures for each of the participants, allowing further future insights. The corpus includes transcripts of 16 groups of either two or three participants playing the murder mystery game within a 90 minute timeslot. Six of these transcripts were annotated using IAT in order to reveal the argumentation structure and to identify the hypotheses participants made. RIP2 is embedded within a wider psychological experiment, with standardised procedures, improved annotation guidelines, and an increased focus on holistic analyses of hypothesis-making and argumentation within the forensic domain.

As such, RIP2 is a resource that allows insight into multiple aspects of research: how participants reason with hypotheses, how group dynamics evolve, and how one might approximate argumentation quality. It contributes a large resource that can be used to investigate human behaviour, as well as train or fine-tune machine learning models in natural language processing (NLP). With such a large resource, we are able to draw conclusions from the data that speak to how humans build and develop hypotheses, paving the way for further research.

2. Related Work

Murder mystery games are designed to be solved through structured logical reasoning, which makes them an accessible method for exploring the construction of arguments. However, how humans succeed at the complex task of reasoning in group settings remains an open question (Eger, 2020). Our primary focus is on hypothesis generation as a key component of the reasoning process, particularly within a collaborative setting. As Schad et al. (2024b) observed within RIP1, argumentative interactions such as arguing, asking, and answering questions may help reduce counterproductive reasoning elicited by confirmation bias, while the process of discarding hypotheses one-by-one might be a particularly effective reasoning strategy. This is in line with experimental work demonstrating that a common mistake in selecting questions intended to inform hypotheses is to treat different answers to the same question similarly, overesti-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5494–5505
11–16 May 2026.

mating the importance of the focal hypotheses over the informativeness of individual subcomponents (Slowiaczek et al., 1992). Applied to our context, this suggests that individual arguments that make a suspect appear more or less likely to be a murderer overall might be wrongly prioritised over relevant sub-questions such as whether they had the means, motive, and opportunity to commit the murder.

While traditionally psychological research has explored whether verification or falsification of hypotheses is more effective, it has been argued that both follow the same underlying mechanism of determining overall probability (Poletiek, 2013), suggesting that confirmation bias can often be merely the unlucky outcome of a valid test strategy (Klayman and Ha, 1987). According to Fischhoff and Beyth-Marom (1983), the main challenges in hypothesis generation are the complexity, ambiguity, and non-exhaustive number of options for potential hypotheses, which can result in untestable hypotheses that are not sufficiently distinct from one another. This links hypothesising to decision-making research more generally, as the process of generating mutually exclusive alternatives is an important yet challenging step that can be influenced by different heuristics and biases (Hämäläinen et al., 2024; Kostiuk et al., 2025).

The process of collaborative accepting or discarding of hypotheses may bring together research from hypothesising in the natural sciences (Zimmerman, 2000) to reasoning processes in police investigations, intelligence analysis, and operational planning (Dhami et al., 2019). As Graesser et al. (2018) note, collaborative problem solving is an important skill in a variety of contexts which, due to its complexity, has rarely been studied directly. Based on the limited empirical evidence available, Sun et al. (2020) validated a model describing collaborative problem solving as relying on the construction of shared knowledge, negotiation, and coordination of strategy, as well as an upholding of team function. For instance, negotiating, as opposed to just sharing ideas, differentiates high and low performing university students (Zhang et al., 2022). The importance of argumentation for hypothesising has been implicitly acknowledged for a while, particularly in science education (Fakhriyah et al., 2021; Walker et al., 2012). There has previously been interest in applying argumentation and NLP methods in forensic contexts (Carnaz et al., 2021; Bjelland and Dahl, 2017), with Siegel et al. (2005) first describing an application for hypothesis generation for intelligence analysis. In recent years, artificial intelligence research has begun to explore automated collaborative decision-making using agent-based systems in the context of board games (Lan et al., 2023), also drawing on murder mystery games to aid agents with conversational dynamics such as turn-taking (Nonomura and Mori, 2025).

Defining what constitutes “good” argumentation is difficult, particularly in informal contexts featuring natural language (Hahn, 2020). The artificial context of a game with a definite conclusion allows for the relevance and consistency of arguments to be measured. With sufficient understanding of the game, factors such as logical cogency, rhetorical effectiveness, and dialectical reasonableness (Wachsmuth et al., 2017) can be compared between groups as the creation of a "ground truth" allows to approximate the elusive notion of objective argument quality (Hoffmann, 2018). At the same time, automated detection of hypothesising (Schad et al., 2024a) may help identify where different argumentation strategies diverge, arriving at conclusions through more or less flawed arguments. A similarly annotated dataset (Hautli-Janisz et al., 2022a) has been used in a recent shared task on argument mining (Ruiz-Dolz et al., 2024). The RIP2 dataset expands the potential scope for argument mining tasks, via the additional annotation of hypotheses.

3. Set Up

Informed by RIP1’s early observations of how individual players’ communicative strategies might have influenced group success, RIP2 sought to develop a larger corpus to investigate more closely which factors might impact group reasoning. To improve the reliability of findings, the corpus was embedded within a larger psychological experiment. The groups playing the murder mystery game were selected from a sample of 189 participants (aged 18-61, 45 male) recruited through the university participant pool, SONA Systems[1], and through advertising in the local area. All selected participants had completed a brief online experiment yielding basic demographic information, as well as measures potentially related to game success, such as reasoning ability and previous experience with murder mystery games. The impact of these will be assessed in future studies. Note, however, that participants were not grouped by reasoning ability. The participants who completed the murder mystery game included 11 males, 28 females, and three non-binary participants. 35 were university students and 11 were non-native speakers of English. Demographic information for the selected participants can be found in our RIP2 repository[2]. Tables 1 and 2 break down the details of the corpus.

We invited 55 participants to take part in the murder mystery study; of these, 50 attended, and 42 are included in this dataset. A preregistration of the

[1] https://www.sona-systems.com

[2] https://github.com/e-schad/RIP2

Shortest transcript:	6801 words
Longest transcript:	17508 words
Average transcript length:	11104 words
All transcripts:	177661 words
Total annotated word count:	80000 words
Minimum number of people:	2
Maximum number of people:	3
Total number of people:	42
Shorted recording:	69 minutes
Longest recording:	100 minutes
Average recording time:	88 minutes
Total recording hours:	23.4 hours

Table 1: Transcript and recording details of corpus.

wider psychological experiment is available on the Open Science Framework[3]. Participants were allocated into groups of three and invited for in-person participation in the game. Due to no-shows and late cancellations, some of the sessions were attended by two participants only, resulting in 7 groups of two and 12 groups of three. Though a total of 19 groups were tested, only 16 recordings were transcribed as some of the groups had to be excluded from the study, two due to prior acquaintance, which might influence group dynamics. The six transcripts that were further analysed are all based on groups of three.

All groups were given the same amount of time to complete the game (90 minutes), disentangling group success from duration, though groups could still opt to end the game early if they were certain about their answer. To allow for the game to be completed within the time limit, the game was slightly shortened. Within RIP1, participants had to solve a code as part of the game, which took the only group to succeed a significant amount of time. Not only did this risk group success primarily hinging on succeeding on one particular sub-task of the game, but code-breaking data also had to be excluded from annotation, as the reasoning behind different attempts was nearly impossible to grasp from the linguistic data. Instead, participants were provided with decoded messages and instructed to ignore all reference to code-breaking in the game instructions. Finally, all evidence was numbered and labelled to allow clear references to the materials in the game. Participants were provided with an inventory of all materials, and the materials were placed in the order indicated by the game.

Participants completed a consent form as well as pre- and post-questionnaires capturing their anticipation of and reflection on the group dynamics and expected success. It was emphasised that participants could also deviate from their groups in their prediction about the solution of the game.

The participants were given access to a whiteboard and markers, paper and pens, snacks and water. An author remained with them throughout to monitor audio recording and manage time, as well as answer any questions. Time reminders were given after 30, 45, 60, 75, 85, and 89 minutes, where applicable. At the end of the game, the group was asked to give their joint solution, describing who was the murderer and a brief summary of their reasoning (apart from RIP2a), before completing the post-questionnaires, receiving the game solution, and a £20 Amazon voucher each.

RIP2	Word Count	Length (mins)	Participants
a	9688	100	2
b	12225	94	3
c	13999	91	3
e	9606	69	2
f	6713	90	2
g	10550	92	3
j	13955	85	3
k	10088	84	3
l	15587	96	3
m	11239	79	3
n	8101	93	2
o	12561	86	2
p	7199	94	2
q	17508	83	3
r	6801	76	3
s	11841	93	3
All	**177661**	**1405**	**42**

Table 2: Statistics for RIP2 corpus including word count, length in minutes, and participants.

3.1. Annotation

Six expert annotators annotated the corpus, including some of the authors. They were paid above living wage and the six transcripts took c. 460 hours to annotate; the decision to annotate six transcripts was made due to the expense of manual annotation. The annotators have, on average, four years of experience. With one exception, all annotators worked on the RIP1 corpus, making them very familiar both with the source material and the format. Improving upon previous processes, all materials that were available to the participants were made available to the annotators in a searchable PDF document, allowing them to correctly attribute reported speech as participants often read aloud from the materials.

We annotated for both "hypothesising" and "conjecturing". Hypothesising is supported by evidence and some epistemic commitment, whereas conjecturing requires less epistemic commitment and little to no evidence. In this way, we are able to capture speculative acts (like guessing) that are

[3] https://osf.io/2e45b/overview?view_only=bf557d4463634c4e89388b2b6b1dc947

RIP2	**a**	b	c	e	**f**	g	**j**	**k**	l	**m**	**n**	o	p	**q**	r	**s**
Answer	Cherie	Joan	Chris	Chris	Cherie	Chris	Cherie	Cherie	Chris	Cherie	Cherie	Chris	Chris	Cherie	Chris	Cherie
Success	**Yes***	No	No	No	**Yes**	No	**Yes***	**Yes**	No	**Yes**	**Yes**	No	No	**Yes**	No	**Yes***

Table 3: All subcorpora of RIP2's final answers. Marked with an asterisk: RIP2a were unable to come to a decision by the end of time, but did both name Cherie as their final answer in their separate post-questionnaires; RIP2j and RIP2s stated that Cherie was the murderer, assisted by Chris. Successful groups are emboldened. In total, eight groups were successful, and eight groups were unsuccessful.

less supported than hypothesising, and differentiate them from hypotheses. Our annotation guidelines included three tests that annotators applied in order to help determine the level of epistemic commitment[4]. We additionally annotated locutions in which participants explicitly or implicitly referred to the means, motive, or opportunity of suspects, as well as whether they made concrete statements about who might have been the murderer. This was included as meta-data and is not added to the IAT-level annotation.

3.1.1. Annotation Framework

Inference Anchoring Theory (IAT) models the ways in which arguments unfold, showing how they are interacted with by participants in a dialogue setting (Reed and Budzynska, 2011; Budzynska et al., 2014b)[5]. Locutions, i.e., unedited utterances from the dialogue participants, make up the dialogue structure. These segmentations of texts are argumentative discourse units (ADUs), i.e., "minimal units of discourse" (Peldszus and Stede, 2013) comparable with elementary discourse units (EDUs), the unit typically used in discourse processing (Seyfried et al., 2024). Propositions are locutions that have been reconstructed to the degree that little context is necessary when reading them separately from the dialogue structure. Thus, they are grammatically complete, with anaphoric expressions resolved where possible. Anchoring the locutions and propositions are illocutionary forces such as "hypothesising" and "asserting", with the class of "default illocuting" capturing cases where there is no clear indication of which speech act is intended, often occurring in question-answering. IAT annotates three argument relations: inference, conflict, and rephrase.

4. The Corpus

The entire corpus of sixteen transcripts and the annotated six groups are available on GitHub, with more materials to be added as they become available. Table 3 breaks down the groups by answer and success.

4.1. Example of Annotated Data

Figure 1 represents the format of the data. Participant 4 hypothesises about a motive for a character, which Participant 5 attacks, providing a reason for their attack. The right-hand side shows the locutions, i.e., what was said by the speaker; the left-hand side represents the propositions which have been reconstructed so that little extra knowledge is necessary to understand what is said. This can be seen with Participant 4's utterance which includes a mention of "him": this is reconstructed to "Nick" in the proposition. The yellow ovals that span locutions and propositions are the illocutionary forces and capture the intentions of the speaker. The purple ovals connecting the locutions represent the dialogue moves made by the speaker. The relations between the propositions (here, shown as "default conflict" and "default inference") represent the argument structure.

4.2. The Game

As in RIP1, participants played the game *Death at the Dive Bar*[6]. Although estimated to take 45 minutes, we found it took groups longer to complete (Schad et al., 2024b). Within RIP2, participants had up to 90 minutes to complete the game. The game includes minimal instructions emphasising that the murderer must have had the motive, means, and opportunity to commit the murder.

The murder victim is Nick, owner of a local bar. The game provides four main suspects: Cherie (the murder victim's wife), Chris (the deputy police officer investigating the crime), Joan (a neighbour involved in esoteric activities in the area), and Donna (a real estate agent and loyal customer). The game's solution identifies Cherie as the murderer, though Chris' involvement is also under investigation. Therefore, participants who mentioned Chris as an accomplice were also considered successful.

[4]The hypothesising guidelines used specifically for this corpus, and the general annotation guidelines, are available on the RIP2 repository.

[5]For IAT diagramming we use OVA+, an online tool developed for the analysis of arguments (Janier et al., 2014).

[6]https://www.huntakiller.com/products/death-at-the-dive-bar-murder-mystery-game

	RIP2b	RIP2g	**RIP2j**	**RIP2k**	RIP2l	**RIP2q**	RIP2
Word Count	12765	10748	14919	*10149*	16431	**18367**	83379
Locution Count	688	856	1237	*571*	**1529**	1265	6146
Illocution Count	1031	1274	1841	*877*	**2246**	1869	9138
Locution-to-word density	*5.63%*	8.11%	8.86%	5.66%	**9.81%**	7.23%	7.69%
Inference-to-locution density	16.28%	16.94%	18.43	**21.89%**	*15.5%*	17%	17.28%
Conflict-to-locution density	**6.69%**	5.26%	4.28%	*2.8%*	5.95%	5.14%	5.14%
Rephrase-to-locution density	*18.75%*	19.04%	21.99%	22.07%	**22.43%**	20.16%	20.96%
MMOM-to-locution density	8.58%	9.35%	**22.07%**	14.01%	19.49%	*7.11%*	9%
MMO deviation from even distribution	**16.11%pt**	*3.92%pt*	13.74%pt	11.85%pt	12.55%pt	15.44%pt	13.43%pt

Table 4: Table of analytics for a representative sample of the RIP2 corpus. Emboldened numbers are the highest across the subcorpora, italicised numbers are the lowest across the subcorpora. Successful groups are emboldened.

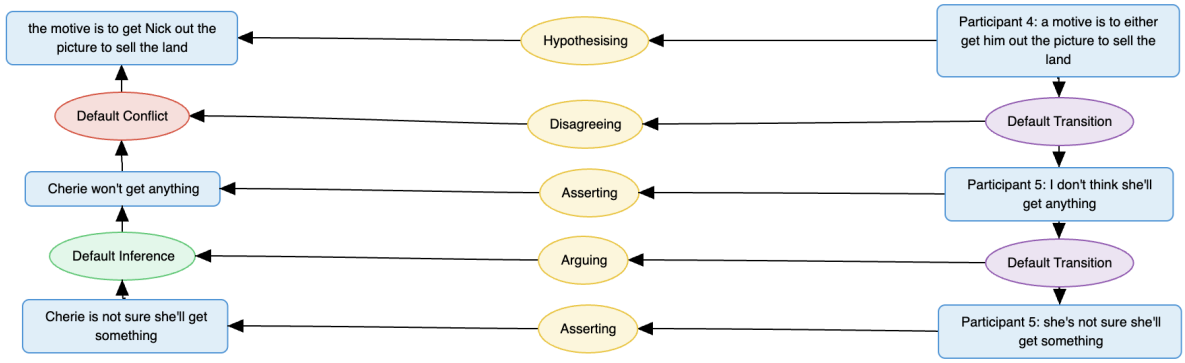

Figure 1: Example of annotated data from RIP2, map ID 33207.

4.3. Inter-Annotator Agreement

Inter-annotator agreement (IAA) is a metric that can indicate data annotation standards. IAA was carried out on 10% of the total annotated corpus; all subcorpora were included as part of the 10%. The annotated corpus achieves a CASS score of 0.38. We use CASS (Duthie et al., 2016) as the metric for our multi-step annotation process, as alternative metrics over-penalise annotation differences arising from early-stage decisions such as including or excluding certain segments from analysis. This agreements score is considered “fair agreement”, as categorised by Landis and Koch (1977). This score may be influenced by the subtle distinctions between “asserting”, “hypothesising”, and “conjecturing”. A small error analysis was undertaken on the IAA subcorpus to investigate the reasons for the lower score. This showed some of the common differences within annotation, which includes different segmentation of reported speech and mixed use of inference and rephrase. The latter can be seen within Figure 2, where one annotator interpreted the text as an argument and the other as elaboration. While complicating the discussion around agreement scores, annotation is a highly complex task and different interpretations of text can be valid (Hautli-Janisz et al., 2022b).

4.4. Analytics

Table 4 gives us an overview of the subcorpora and the argumentation density. Table 5 shows the illocutionary acts used within the subcorpora. Most of these are commonly used in IAT; "asserting" is a default assertion, "arguing", "disagreeing", and "restating" are used to anchor instances of inferences, conflicts, or rephrases respectively, "default illocuting" is used when participants answer questions, "questioning" is used for open or closed questions, and "challenging" involves eliciting reasoning from another participant. "Hypothesising" and "conjecturing" were also specifically analysed for this corpus, describing potential events that are either supported by evidence or are more speculative in nature. Both tables additionally give a total value representing the corpus as a whole. If the group was successful in their hypothesising, we mark this

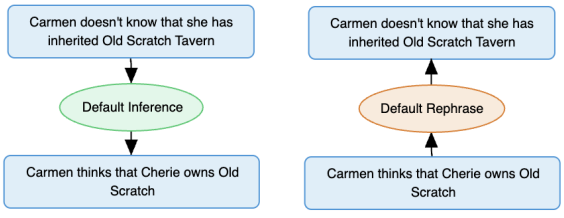

Figure 2: Example where elaboration or support can be intepreted.

by emboldening their subcorpus name in all tables.

Means, motive, opportunity, and murderer (MMOM) were labels that annotators ascribed to locutions in the transcript whenever these topics were considered, as discussed in Section 3.1. Overall, RIP2j had the highest proportion of MMOM references. In other words, this group's discussion was particularly rich in references to murder-relevant discussions. We also calculated the deviation from even distribution as a measure of how balanced each group's discussion was in terms of paying equal attention to means, motive, and opportunity (MMO), as per the game's instructions. This is based on the average of the groups' absolute deviations from an ideal relative proportion of 1/3 for each of the three topics of MMO, as participants were explicitly told within the instructions that these three aspects would help them solve the mystery. RIP2g, while having quite low locution-to-word density, i.e., a small proportion of argumentative content to total speech, was the most even in terms of the discussion out of the subcorpora.

4.4.1. Combative and Cooperative

Table 4 shows that RIP2k had the highest inference density and the lowest conflict density at 2.8%. RIP2b, by comparison, had the highest conflict-to-locution density nearing 7%.

As shown in Table 5, RIP2l's participants agreed a lot (13.18%), particularly in contrast to RIP2k, whose participants agreed the least (4.33%). RIP2l similarly had the highest amount of challenging (0.80%), as well as the highest amount of disagreeing (5.12%). This mix of high agreement, high disagreement, and a lot of challenging shows the members of RIP2l to be very engaged in the discussion. While RIP2k has the lowest agreement, it also had the lowest disagreement (2.05%). RIP2j also had little disagreement (3.31%), but relatively high agreement (12.87%).

4.4.2. Support and Elaboration

Groups RIP2l and RIP2b were the least inferentially dense of the subcorpora, at 15.5% and 16.28%. Table 4 additionally shows that RIP2l had the highest rephrase density, although overall all corpora had a high density of rephrases, suggesting that elaboration was common among participants. RIP2b had both the lowest rephrase-to-locution density and lowest amount of restating. Taking into consideration that the subcorpus was one of the lowest for inference and arguing as well, RIP2b did not support or extrapolate their utterances as much as the other subcorpora. RIP2k had the highest inference-to-locution density, with a difference of over six percentage points from the lowest subcorpus. In line with RIP2k's high inferential density, in Table 5 it had the highest percentage of arguing among the subcorpora. RIP2j had high amounts of restating and the second highest amount of arguing, suggesting they both supported and elaborated upon their utterances.

4.4.3. Asking and Answering

Groups RIP2b, RIP2k, and RIP2q asked and answered the most questions out of the six subcorpora; see Table 5. Of the three, RIP2b asked the most questions, but RIP2q answered the most. Proportionally, groups RIPj and RIP2q answered more questions than the overall corpus (respectively 2.06% and 2.14% of questions were left unanswered, with 2.63% overall) whereas RIP2l and RIP2b left the most unanswered (3.20% and 3.21%). RIP2j asked and answered the least questions out of the six teams.

5. Insights into Hypothesising

Table 6 describes the percentage of the different relations (inference, rephrase, and conflict) that connect to instances of hypothesising. An infer-

	RIP2b	RIP2g	**RIP2j**	**RIP2k**	RIP2l	**RIP2q**	RIP2
Agreeing	9.70%	7.54%	12.87%	*4.33%*	**13.18%**	9.20%	10.28%
Arguing	10.38%	11.30%	12.44%	**14.03%**	*10.24%*	11.29%	11.42%
Asserting	48.88%	**54.47%**	48.18%	54.05%	*45.81%*	48.53%	49.19%
Challenging	0.29%	0.55%	*0.11%*	0.46%	**0.80%**	0.59%	0.49%
Conjecturing	**2.04%**	1.41%	2.01%	*0.80%*	1.29%	1.98%	1.63%
Default Illocuting	3.88%	2.51%	*2.44%*	3.76%	3.21%	**4.39%**	3.33%
Disagreeing	4.66%	4.08%	3.31%	*2.05%*	**5.12%**	3.91%	4.02%
Hypothesising	**4.36%**	2.90%	*1.74%*	3.31%	2.00%	4.23%	2.92%
Questioning	**7.08%**	5.02%	*4.51%*	6.61%	6.41%	6.53%	5.95%
Restating	*8.73%*	10.20%	**12.38%**	10.60%	11.80%	9.36%	10.74%

Table 5: Percentage of illocutionary acts for the RIP2 corpus. Emboldened numbers are the highest across the subcorpora, italicised numbers are the lowest across the subcorpora. Illocutionary acts that are less than one percent of the corpus are omitted. Successful groups are emboldened.

	RIP2b	RIP2g	**RIP2j**	**RIP2k**	RIP2l	**RIP2q**	RIP2
% of Inferences to Hypotheses:	19.64	11.72	*13.60*	15.20	18.99	**26.05**	17.89
% of Conflicts to Hypotheses:	**13.04**	*4.44*	5.66	6.25	8.79	12.31	8.86
% of Rephrases to Hypotheses:	4.65	*1.23*	2.94	**7.14**	4.66	6.27	4.43

Table 6: Table of Hypothesising Analytics for the RIP2 corpus: percentages of relations that connect to hypotheses. The relation to hypothesis is divided by total relations. Emboldened numbers are the highest across the subcorpora, italicised numbers are the lowest across the subcorpora.

	RIP2b	RIP2g	**RIP2j**	**RIP2k**	RIP2l	**RIP2q**
Cherie	**57.69%**	*0%*	35%	47.06%	31.25%	30.43%
Chris	7.69%	**83.33%**	30%	*0%*	18.75%	30.43%
Joan	7.69%	*0%*	5%	23.53%	**50%**	8.70%
Donna	26.92%	16.67%	30%	29.41%	*0%*	**30.43%**
Deviation from even distribution	14.72%pt	**17.91%pt**	*9.38%pt*	11.64%pt	16.11%pt	10.44%pt

Table 7: Percentage of hypotheses across the annotated RIP2 corpus focused on suspects, including deviation from even distribution. Emboldened numbers are the highest across the subcorpora, italicised numbers are the lowest across the subcorpora.

ence connecting to a hypothesis means that acts of hypothesising were supported through evidence or reasoning. On the other hand, a conflict to a hypothesis shows that a hypothesis was attacked or rejected. Rephrases to hypotheses indicate that a similar hypothesis was made, potentially extending or narrowing down an existing hypothesis. RIP2g supported and conflicted with hypotheses the least, as well as elaborating on them the least. A quarter (26.05%) of all of RIP2q's inferences supported hypotheses. Alongisde RIP2q, RIP2b (19.64%) and RIP2l (18.99%) supported their hypotheses the most; they also conflicted with them the most. The analytics show a trend of hypotheses being supported more than they are conflicted with or elaborated upon. 28 conflicts, of 316 in total, attacked hypotheses (of which there are a total of 267 hypotheses). As such, only 8.86% of conflicts are conflicts to hypotheses, showing that hypotheses might be generally discarded indirectly when moving on to a new hypothesis. So, when the team finds evidence that the culprit cannot be Donna, instead of directly rejecting their previous hypothesis that "the murderer is Donna", they might be more likely to move on to a new hypothesis.

Rephrasing hypotheses is even less common than supporting or contrasting, though there are overall more rephrases than conflicts (around 4% of the corpus is made up of disagreement, versus over 10% for restating). Hypotheses are shown here to be relatively unlikely to be iterated upon, which may reflect the restricted game environment which allows only a few options and suspects.

5.1. Suspect Hypotheses

Table 7 shows how discussion over suspects aligned with their eventual accused suspect. RIP2b gave Joan as their final answer, yet the majority of their hypotheses were about Cherie. Looking at the transcript, it is clear that the group only late on in their discussion came to suspect Joan when there was little time left to discuss her. Before that point, they had suspected Cherie.

We can see how close groups were to an even distribution, a lower number being more desirable as it shows that their discussions were not only focused on a single suspect, but more evenly distributed. Overall, the unsuccessful groups had an average deviation of 16.25 percentage points, in contrast to the successful groups with an average deviation of 10.49 percentage points. Successful groups, therefore, were more evenly distributed in their hypothesising about the four suspects. Focusing only on one suspect can lead to confirmation bias, which Analysis of Competing Hypotheses (ACH) matrices (Heuer, 1999) are used to help combat, and our findings here indeed suggest that a more even distribution of suspect hypothesising might help in terms of success.

5.2. Successful vs Unsuccessful

Table 8 indicates that successful groups are more likely to hypothesise about suspects than unsuccessful groups, suggesting that they made more relevant and on-topic hypotheses. Following this, Table 9 shows that successful groups also supported suspect hypotheses more. They were 20 percentage points more likely to support hypotheses and around 13 percentage points more likely to rephrase or conflict with them. Overall, successful groups interacted more with their suspect hypotheses than the unsuccessful groups.

	% of Suspect Hypothesising to Total Hypothesising
Unsuccessful	42.52%
Successful	59.29%

Table 8: Total percentage of hypotheses across the annotated RIP2 corpus focused on suspects.

While overall there were only small differences in how successful and unsuccessful groups argued with all instances of hypothesising, shown by Table 10, there was a larger difference, between 12 and 22 percentage points, when specifically looking at suspect hypotheses: see Table 9.

	RA	CA	MA
Unsuccessful	53.64%	37.36%	14.33%
Successful	75.70%	50.00%	26.34%

Table 9: Total percentage of inferences (RAs), conflicts (CAs), and rephrases (MAs) linked to suspects across the annotated RIP2 corpus.

Table 10 additionally shows that successful groups agreed more (with a difference of 3 percentage points) and disagreed more (with a difference of 9 percentage points) with all instances of hypothesising. The data thus far suggests that interacting more with hypotheses, particularly on-topic ones, might be linked with a successful outcome.

	Unsuccessful	Successful
Agreeing	3.46%	6.71%
Disagreeing	5.71%	13.64%
Inference	17.46%	18.83%
Conflict	10.46%	9.76%
Rephrase	4.95%	6.65%

Table 10: Agreeing, disagreeing, and relations to all instances of hypothesising across the annotated RIP2 corpus.

6. Discussion

Out of 16 groups, eight were successful and eight were unsuccessful. Two groups put forward Cherie being the murderer but with Chris as an assistant and only one team nominated a different culprit from either Cherie or Chris. The 80k annotated corpus is made up of 2.92% hypothesising and 1.63% conjecturing illocutionary types, with an inference-to-word density of 1.33% and a conflict-to-word density of 0.40%.

Successful groups interacted more with on-topic hypotheses than the unsuccessful groups did, despite there being no difference between the two groupings' use of relations to all hypotheses. The successful groups made more balanced on-topic hypotheses and interacted more with them; hypotheses were not used more by successful groups, suggesting that the presence of hypotheses are not indicators of success, but that the degree of interaction with them might be. Successful groups had a more even distribution of hypotheses about suspects, with a difference of six percentage points. In intelligence work ACH matrices are used as a way to prevent bias (such as confirmation bias) by encouraging users to look for competing hypotheses and evidence for them. These results suggest that a more even distribution of hypotheses over suspects may be helpful for success.

Where Schad et al. (2024b) explored how the argumentative strategies of individual group members might have influenced group success, here we focus on argumentative interaction at the group level. The three successful groups illustrate the importance of shared knowledge, strategy, and teamwork in the process (Sun et al., 2020). Though RIP2k is the shortest annotated subcorpus, it features the highest density of inference and the highest proportion of rephrased hypotheses, while being low in conflict, disagreeing, and conjecturing. This suggests that RIP2k might be characterised by high reasoning quality and careful treatment of hypotheses, with little need for disagreement. RIP2j, on the other hand, appears to be the most strategic, as the group was the most frequent in discussing means, motive, and opportunity, which directly related to group success as per the game instructions. While RIP2j does not feature much diversity of illocutionary forces, including the lowest proportion of hypothesising, their hypotheses are the most balanced between different suspects, further facilitating their success. RIP2q is based on the longest transcript, and this group appears quite cohesive. Even though the topics of MMOM are least discussed in this group, it has the highest proportion of question answering and inferences to hypotheses.

Across analyses, RIPl also appears very engaged. This is the group with the highest proportion of argumentative content, which also uses the highest percentage of rephrases and the most agreeing, disagreeing, and challenging. Despite the importance of negotiation for the generation of well-formed hypotheses, RIP2l was unsuccessful, which might suggest a failure to produce good reasoning despite good discussions. On the other hand, this might also be the result of the simplified distinction between success and failure in this paper, although several groups, including RIPl, got "pretty close" (to quote from the transcript). Here, future analyses might help disentangle to what extent group success is directly dependent on good reasoning, or to what extent chance plays a role in choosing the eventual main suspect. The other two unsuccessful

groups, RIPb and RIPg, can be more clearly distinguished from the successful groups. RIPb used the most questioning, hypothesising, and conjecturing, but also had the most conflict, both overall and in terms of disagreeing with hypotheses. This group was also the least balanced in their reasoning about means, motive, and opportunity, suggesting that RIPb was strong in building hypotheses but potentially misguided in focus. Similarly, RIPg, which was the most balanced in terms of MMO, was also the most skewed when it came to hypothesising about suspects, with the vast majority of their hypotheses focusing on only one suspect, and two suspects being completely disregarded. This group also engaged least with hypotheses in terms of inferences, conflicts, and rephrases, suggesting a general lack of focus.

7. Future Work

With 80k words annotated, RIP2 offers plenty of opportunity for future research. 10 corpora still remain to be explored, with all transcripts available alongside the annotated data on GitHub. Since all groups were asked to clearly state their suspicion and describe their reasoning at the end of the game, separate analyses of these reasoning summaries are also possible. Once available, participant background information and pre- and post-game questionnaires can also be drawn on. In particular, the individual answers from participants which sometimes differed from group decisions remain an avenue to explore in group dynamics: how individual participants ideas may be spurned for the overall group decision.

Considering the groups played a murder mystery game with a set solution, we plan on providing further annotations of the argumentation underlying the game's solution, so that future work will be able to compare different groups in terms of their reasoning quality. Since RIP2 is rich in annotations of hypotheses, including those specifically related to the means, motive, opportunity, and murderer, analyses of the emergence of common ground are also possible (Stalnaker, 2002), as are studies of how group success develops over the time. Finally, this study provides further potential for hypothesis mining (Schad et al., 2024a) and computer-assisted applications for collaborative decision-making (Di Maro et al., 2025).

8. Conclusion

RIP2 is the largest existing hypothesising corpus, offering 177k words of collaborative problem-solving, with 80k words annotated using IAT, an argumentation framework that annotates dialogue acts, propositions, and illocutionary acts. It is a five-fold data increase from the other existing corpus on hypothesising, allowing for better analysis and more insights. Within this work we looked at potential factors that influence success, and characterised the subcorpora as well as the corpus as a whole. We discussed how participants use hypotheses, as well as the argumentation interacting with them. We find that successful groups were more likely to hypothesise about suspects than unsuccessful groups by 17 percentage points. Successful groups were also more likely to support, conflict, and rephrase their hypotheses about suspects, despite there being little difference in how successful and unsuccessful groups used these relations for all hypotheses, suggesting that on-topic hypothesis discussions can be particularly important.

RIP2 offers new ways into looking at how human participants collaborate and work together, and their strategies for achieving their aims. By investigating how participants use hypotheses, the corpus offers exciting avenues of exploration and potential application. Exploring hypotheses in an argumentation framework allows for a broad view of what happens in the dialogue, and gives us insight such as that successful groups hypothesised more evenly across the four suspects, rather than only focusing on a single suspect, which corresponds to the use of ACH matrices that explicitly encourage users to consider competing hypotheses. Embedding corpus analyses within experimental research allows us to pave the way to more reliable insights on what might make humans successful at reasoning with hypotheses.

9. Limitations

Although the annotated corpus is almost twice the size of RIP1, findings from this study still remain largely exploratory. Continuing analysis will increase sample sizes to also allow for significance testing. For example, full analysis of the entire corpus, beyond the six groups already analysed, can elucidate to what extent the descriptive observations in this paper are indeed reliable. The corpus can be further replicated and extended by following the procedures outlined in this study. Following participant drop-outs, the number of participants in each group varied from two to three, which risks inconsistencies in comparing different groups. However, the groups selected for annotation all included the same number of participants. Finally, the RIP corpora employ a single, fixed-solution murder mystery game setting to model forensic decision-making, which might limit the generalisability of findings across domains.

10. Acknowledgements

We thank the annotation team for their time and effort in the creation of RIP2. This research is supported in part by Volkswagen Stiftung under grant Az. 98 543; and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

11. Ethical considerations

Ethics approval was granted by the University of Dundee. All participants signed a consent form prior to taking part in the study. Consent to record the murder mystery game was additionally obtained. All transcripts were handled by a trusted company and pseudonymised before given to annotators. Participants could withdraw their consent up to this point; none did so. All participants completing the murder mystery game were reimbursed with £20 Amazon vouchers, regardless of their success in the game.

12. Bibliographical References

Heidi Fischer Bjelland and Johanne Yttri Dahl. 2017. Exploring criminal investigation practices the benefits of analysing police-generated investigation data. *European Journal of Policing Studies*.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014a. Towards argument mining from dialogue. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 185–196. IOS Press.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014b. Towards argument mining from dialogue. In *Computational Models of Argument*, pages 185–196. IOS Press.

Gonçalo Carnaz, Mário Antunes, and Vitor Beires Nogueira. 2021. An annotated corpus of crime-related portuguese documents for nlp and machine learning processing. *Data*, 6(7):71.

Mandeep K Dhami, Ian K Belton, and David R Mandel. 2019. The "analysis of competing hypotheses" in intelligence analysis. *Applied Cognitive Psychology*, 33(6):1080–1090.

Maria Di Maro, Martina Di Bratto, Sabrina Mennella, Antonio Origlia, and Francesco Cutugno. 2025. Argumentation in recommender dialogue agents (arda): An unexpected journey from pragmatics to conversational agents. *Open Linguistics*, 11(1):20250052.

Rory Duthie, John Lawrence, Katarzyna Budzynska, and Chris Reed. 2016. The cass technique for evaluating the performance of argument mining. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 40–49.

Markus Eger. 2020. Murder mysteries: the white whale of narrative generation? In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, pages 210–216.

Fina Fakhriyah, Ani Rusilowati, Endang Susilaningsih, et al. 2021. Argument-driven inquiry learning model: A systematic review. *International Journal of Research in Education and Science*, 7(3):767–784.

Baruch Fischhoff and Ruth Beyth-Marom. 1983. Hypothesis evaluation from a bayesian perspective. *Psychological review*, 90(3):239.

Arthur C. Graesser, Stephen M. Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W. Foltz, and Friedrich W. Hesse. 2018. Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19:59–92.

Ulrike Hahn. 2020. Argument quality in real world argumentation. *Trends in Cognitive Sciences*, 24(5):363–374.

Raimo P Hämäläinen, Tuomas J Lahtinen, and Kai Virtanen. 2024. Generating policy alternatives for decision making: A process model, behavioural issues, and an experiment. *EURO Journal on Decision Processes*, 12:100050.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022a. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Annette Hautli-Janisz, Ella Schad, and Chris Reed. 2022b. Disagreement space in argument analysis. In *Proceedings of the 1st Workshop on*

Perspectivist Approaches to NLP @LREC2022, pages 1–9, Marseille, France. European Language Resources Association.

Richards J Heuer. 1999. *Psychology of intelligence analysis*. Center for the Study of Intelligence.

Michael HG Hoffmann. 2018. The elusive notion of "argument quality". *Argumentation*, 32(2):213–240.

M. Janier, J. Lawrence, and C Reed. 2014. OVA+: An argument analysis interface. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 463–464, Pitlochry. IOS Press.

Joshua Klayman and Young-won Ha. 1987. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2):211–228.

Yevhen Kostiuk, Clara Seyfried, and Chris Reed. 2025. Automating alternative generation in decision-making. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2023. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. *arXiv preprint arXiv:2310.14985*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Ryota Nonomura and Hiroki Mori. 2025. Who speaks next? multi-party ai discussion leveraging the systematics of turn-taking in murder mystery games. *Frontiers in Artificial Intelligence*, 8:1582287.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Fenna H Poletiek. 2013. *Hypothesis-testing behaviour*. Psychology Press.

Chris Reed and Katarzyna Budzynska. 2011. How dialogues create arguments. In *Proceedings of the 7th Conference of the International Society for the Study of Argumentation (ISSA)*, pages 1633–1645. SicSat Amsterdam.

Ramon Ruiz-Dolz, John Lawrence, Ella Schad, and Chris Reed. 2024. Overview of dialam-2024: Argument mining in natural language dialogues. In *11th Workshop on Argument Mining, ArgMining 2024*, pages 83–92. Association for Computational Linguistics (ACL).

Ella Schad, Kamila Górska, Eimear Maguire, Ramon Ruiz-Dolz, Melvin Abraham, John Lawrence, and Jacky Visser. 2024a. Annotating and mining hypotheses in argumentation. In *The 10th International Conference on Computational Models of Argument*, pages 253–264. IOS Press.

Ella Schad, Jacky Visser, and Chris Reed. 2024b. The rip corpus of collaborative hypothesis-making. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16047–16057.

Clara Seyfried, Chris Reed, and Yuki Kamide. 2024. Defining argumentative discourse units as clauses: Psycholinguistic evidence. In *Proceedings of COMMA 2024*, pages 277–288. IOS Press BV.

Nick Siegel, Blake Shepard, John Cabral, and Michael Witbrock. 2005. Hypothesis generation and evidence assembly for intelligence analysis: Cycorp's noöscape application. In *Proceedings of the 2005 International Conference on Intelligence Analysis (IA 2005), McLean, VA, USA*.

Louisa M. Slowiaczek, Joshua Klayman, Steven J. Sherman, and Richard B. Skov. 1992. Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition*, 20(4):392–405.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Joi Phelps Walker, Victor Sampson, Jonathon Grooms, Brittany Anderson, and Carol O Zimmerman. 2012. Argument-driven inquiry in undergraduate chemistry labs: The impact on students' conceptual understanding, argument skills, and attitudes toward science. *Journal of college science teaching*, 41(4):74–81.

Si Zhang, Qianqian Gao, Mengyu Sun, Zhihui Cai, Honghui Li, Yanling Tang, and Qingtang Liu. 2022. Understanding student teachers' collaborative problem solving: Insights from an epistemic network analysis (ena). *Computers & Education*, 183:104485.

Corinne Zimmerman. 2000. The development of scientific reasoning skills. *Developmental Review*, 20(1):99–149.