The *RIP* Corpus of Collaborative Hypothesis-Making

Ella Schad, Jacky Visser, Chris Reed
Centre for Argument Technology
University of Dundee, Dundee DD1 4HN, UK
{firstname}@arg.tech

Abstract

The dearth of literature combining hypothesis-making and collaborative problem solving presents a problem in the investigation into how hypotheses are generated in group environments. A new dataset, the Resolving Investigative hyPotheses (RIP) corpus, is introduced to address this issue. The corpus uses the fictionalised environment of a murder investigation game. An artificial environment restricts the number of possible hypotheses compared to real-world situations, allowing a deeper dive into the data. In three groups of three, participants collaborated to solve the mystery: two groups came to the wrong conclusion in different ways, and one succeeded in solving the game. RIP is a 49k-word dialogical corpus, consisting of three sub-corpora, annotated for argumentation and discourse structure on the basis of Inference Anchoring Theory. The corpus shows the emergent roles individuals took on and the strategies the groups employed, showing what can be gained through a deeper exploration of this domain. The corpus bridges the gap between these two areas – hypothesis generation and collaborative problem solving – by using an environment rich with potential for hypothesising within a highly collaborative space.

Keywords: argumentation, collaborative problem solving, corpus, hypotheses

1. Introduction

Hypotheses are a key element in problem solving and investigation, and there is a body of work on their formulation and evaluation (Gettys and Fisher, 1979; Benjamin, 2019). There is also a growing interest in collaborative problem solving, particularly within a learning environment (Cesareni et al., 2016). However, there remains a gulf between these two fields; a synthetic environment allows the future possibility of investigating real-use cases. This paper, therefore, introduces a new dataset, the Resolving Investigative hyPotheses (RIP) corpus, which comprises of 49,000 words annotated in the Inference Anchoring Theory framework (Reed and Budzynska, 2011; Budzynska et al., 2014) and three groups of three participants each.

Investigating the formation of hypotheses in mock-serious and finite domains creates a stable and helpfully restrictive environment. Real use-cases, such as police investigations or intelligence, do not have a single conclusion to a question or hypothesis, but multiple. The use of a game – a simulated environment – is useful in the exploration of hypotheses. Instead of an unlimited amount of hypotheses being available, this insurmountable list is whittled down to the few that a game developer has created. Participants may still go outside of those stated boundaries (an example of this is discussed in section 4), but this is highlighted as an anomaly and the reasoning for it can be understood within its context. Using a murder mystery game as the domain forcibly limits the available hypotheses that participants can choose from and thus allows greater levels of comparison between the participating groups. The formation and evaluation of hypotheses has been explored (Gettys and Fisher, 1979; Benjamin, 2019), as well as the role of collaborative working in problem solving, particularly within a learning environment (Cesareni et al., 2016). The corpus is an investigation into the field of collaborative hypothesis-making and the kind of arguments that are employed in this area. This dataset contributes to a deeper investigation into hypothesis generation, particularly in investigative environments. This dataset shows the initial results of exploring individual roles and group strategies that are present in the data: section 4. This paper is a first exploration into these two domains using a game to approximate real use-cases. The corpus is made accessible online.[1]

2. Related Work

The literature on hypotheses stretches from philosophy of science (Slowiaczek et al., 1992; Zimmerman, 2000) and the psychology of hypothesising (Klayman and Ha, 1987; Bassok and Trope, 1984), to their probabilistic evaluation (Fischhoff and Beyth-Marom, 1983). Hypothesis generation can be both an individual or collaborative process, and thus investigating the literature naturally leads to collaborative problem solving. Participants within the RIP corpus create hypotheses both individually and collaboratively, since murder mystery games are designed to be solved in a group setting. Collaborative problem solving (Koutsombogera and

[1] http://corpora.aifdb.org/mm123

LREC-COLING 2024, pages 16047–16057

Vogel, 2018) has a strong psychological foundation (Slavin, 2011; Gigone and Hastie, 1997). Collaborative work, however, forms a more general class and according to Graesser et al. (2018), differs from collaborative problem solving; it is a very structured type of problem solving, including a differentiation of roles and team interdependency. In other works collaborative and cooperative problem solving are differentiated; the latter being a matter of a division of labour, and the former a mutual attempt at the problem (Roschelle and Teasley, 1995).

There is a lack of literature exploring the use of hypothesis generation within police investigations or intelligence analysis, although some literature is interested in using tools to automatically generate hypotheses (Siegel et al., 2005). Carnaz et al. (2021) builds a corpus of crime-related Portuguese documents in order to employ NLP methods, and other areas attempt to demystify and expand the potential of police-generated data (Bjelland and Dahl, 2017). Bex (2011) explores the reasoning of evidence and the use of narratives in regards to criminal cases.

This type of task, the working towards a conclusion with a finite amount of time and resources, also has links to superforecasting projects (Tetlock and Gardner, 2016; Katsagounos et al., 2021). The environment also uses artificiality – posing (as-of-yet) unanswerable question about possible future events and asking participants to answer (i.e., hypothesise) within the constraints. It was also explicitly set up so the answers were to be revealed in the near future, allowing exploration into participants' reasoning and processes; having a concrete answer is therefore a crucial part to the investigation of hypothesis generation.

Differing from the Penn Discourse Treebank corpus (Prasad et al., 2019), this corpus is a more compact and specific one. Through RIP's fictitious murder mystery domain, an exploration into the generation of hypotheses (in a deliberately limited way) is made possible and desirable.

3. The Corpus

3.1. The Game

The game used, *Death at the Dive Bar*[2], published by Hunt a Killer, was marked as "Easy" and was estimated to take 45 minutes; all participants were aware of this fact. The use of a game environment is two-fold: to simplify and restrict the number of possible hypotheses, but also to approximate real use-cases. Real investigations are not be a perfect match to a murder mystery game, but there are some inherent similarities. In discussing the data here, the solution inevitably will be spoiled. The game had one correct answer which was given at the end of the game. The game included evidence, fact sheets, and physical objects that were relevant to the solving of the game. It had a specific story and the groups were given a set of instructions which set the scene and set up the game; for instance, these instructions noted the importance of revealing motive, means, and opportunity, as well as there being four suspects (Cherie, Chris, Joan, and Donna). A brief outline of the game is as follows: Nick, the victim, was murdered by his wife Cherie who was having an affair with Chris, a police deputy. Chris and Donna had alibis that relied on one another, Cherie had an alibi only for the hour before the time of death, and Joan had another character as a witness for her alibi. These four are the official in-game suspects; Group 3 also included, and later accused, Carmen, the character who instigated the investigation for the players. There was also some code-cracking involved in the unravelling of the game.

Gameplay corpora often centre around group interaction and competition (Hung and Chittaranjan, 2010), how players themselves conceive of the game (Shaker et al., 2011), and of strategy (Lewis et al., 2011). Guhe and Lascarides (2014) test strategies employed in the game, *The Settlers of Catan*, to improve an autonomous agent's gameplay performance. This paper will use the domain that the game provides to investigate both group and individual dynamics and in order to restrict the number of possible hypotheses.

3.2. The Process

Three groups of three participants were tasked with playing the game, *Death at the Dive Bar*. The three groups were given the same information and the same task. No roles were designated and no instructions outside of the game were given. The instructions within the game were taken somewhat fluidly. Two groups deviated from the specified instructions: Group 1 chose two suspects and Group 2 chose outside of the clearly defined four suspects. All groups were given paper to write on, but the third group additionally utilised a large touchscreen in the room to record timelines and motives.

All groups differed in their final accusation. Group 1 thought it was a joint effort of Cherie (the culprit) and the man she was having an affair with, Chris, and that of the two, Chris was the most likely to have done the killing. Group 2 put forward Carmen, a non-suspect, as the murderer. Group 3 correctly put forward Cherie as the murderer. The groups also spent differing amounts of time playing the game: Group 1 spent an hour and a half on the task; Group 2 spent an hour and forty minutes; and Group 3 spent two hours and forty minutes.

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

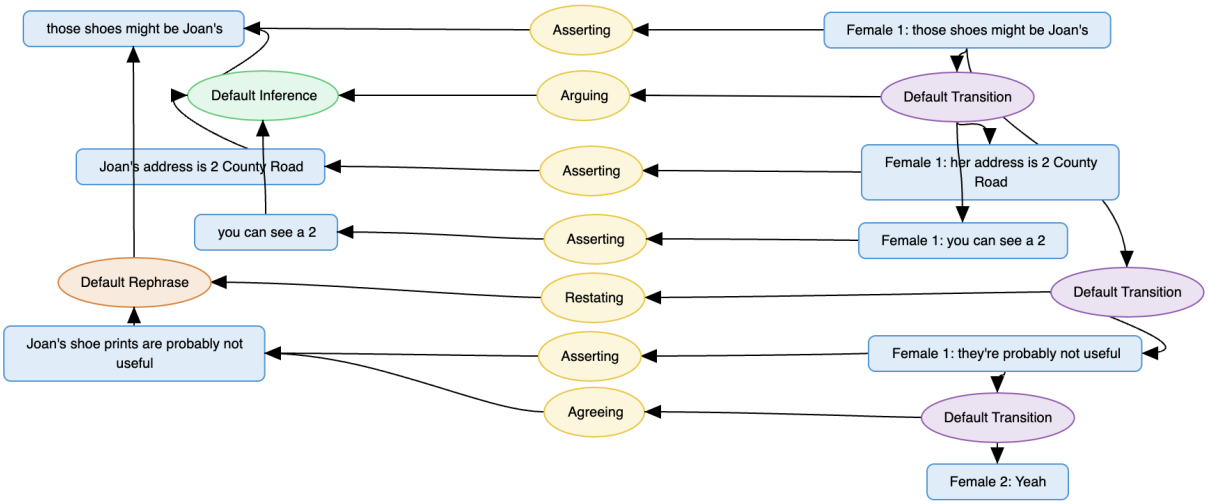

Figure 1: Example of IAT annotation: rectangular blue boxes indicating propositions (left-hand side) and locutions (right-hand side), yellow ovals for illocutionary connections, purple ovals for discourse transitions, the green oval for inference relation, the orange oval for rephrase relation

3.3. Annotation

3.3.1. The Annotation Process

Inference Anchoring Theory (IAT) models arguments and the ways in which they interact in dialogical environments. This includes how arguments are supported and attacked and how they unfold in discourse. The framework uncovers both the propositional and dialogue structure of the discourse, with the latter anchoring the former.[3] The right-hand side captures the dialogue structure in the form of locutions, the utterances of interlocutors that remain unedited. These segmentations of texts are argumentative discourse units (ADUs). ADUs are "minimal units of discourse", as introduced and defined in Peldszus and Stede's (2013) survey on argument mining. They are non-overlapping spans of text that have discrete argumentative function and are comparable with elementary discourse units (EDUs), the unit typically used in discourse processing.

The propositional structure is shown on the left-hand side. These propositions are reconstructed in regards to grammar and resolving anaphoric and elliptical expressions. Propositional content must be reconstructed to the degree that little further explanation of background knowledge is needed. Each ADU must be a parsable sentence and therefore is typically reconstructed with a verb and a noun.

IAT allows close analysis of the argument relations in dialogue, capturing interlocutors' dialogical interaction. There are three main relations in IAT: support (shortened to RA), attack (shortened to CA), and rephrase (shortened to MA). In Figure 1, Default Inference (support) and Default Rephrase are used. They are anchored through the illocutionary forces of Arguing and Restating respectively. Locutions, the right-hand side of the graph, anchor propositions with Asserting. The illocutionary forces show how the dialogue is being used and the speaker's likely intentions – locutions and propositions are anchored through an Asserting in this case. If the interlocutor were to ask a question, however, this would be reflected in the illocutionary force shown: Pure Questioning, Assertive Questioning, or Rhetorical Questioning. The purple Default Transitions capture the functional flow of the conversation (Wells and Reed, 2012); they show temporality only in that a reply or continuation of conversation can only follow from subsequent locutions. They show the following, or breaking, of dialogue games. If two interlocutors discuss the weather and then one of the speakers starts discussing their washing machine, this break in the dialogue game would be reflected in the lack of a Default Transition. The interlocutor may then explain their break of the game with how the weather affects their drying of the washing, thereby connecting the conversations together. This would necessitate a long-distance Default Transition: long-distance in that it connects locutions which are not temporally adjacent.

Figure 1 is an example from the RIP corpus, which illustrates a linked argument, an Agreeing, and a Default Rephrase. Focusing on the left-hand side, the conclusion of the linked argument ("those shoes might be Joan's") is shown through the direction of the arrows: arrows point *from* the premise *to* the conclusion. A linked argument is one where

[3]For IAT diagramming we use OVA+, an online tool developed for the analysis of arguments (Janier et al., 2014). The IAT framework and its OVA tool have been used for more than 2.5 million words of analysed argumentation.

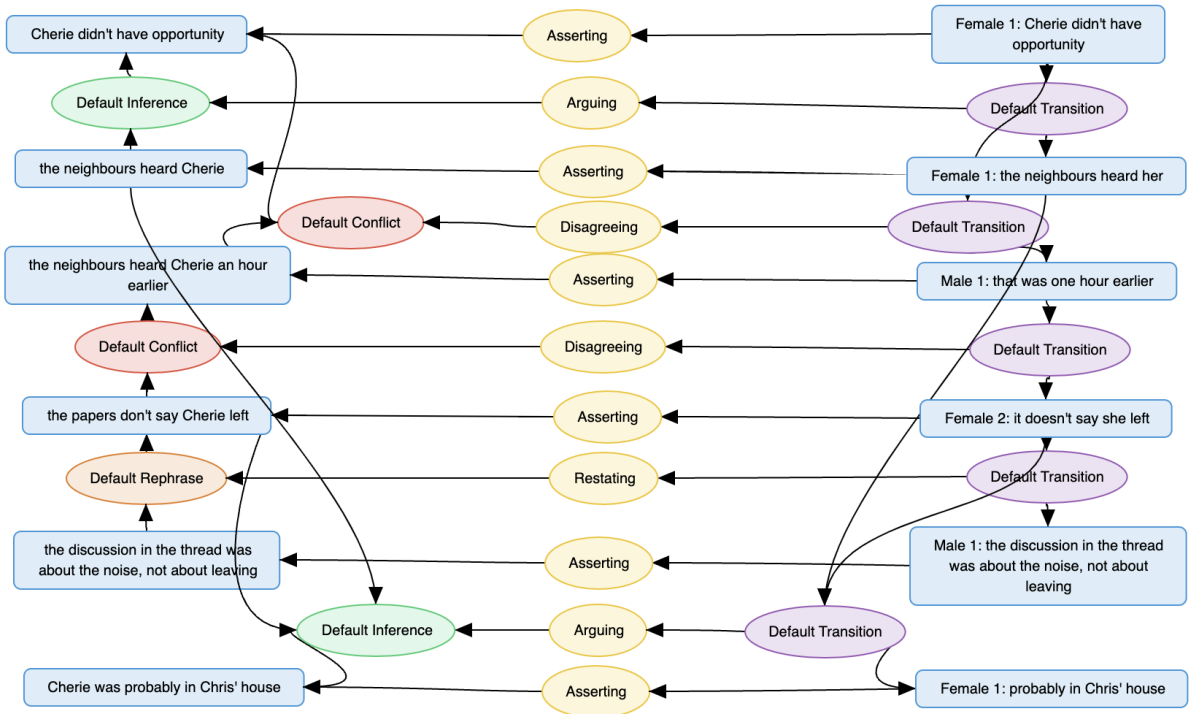

Figure 2: Example of Annotated Data: rectangular blue boxes indicating propositions and locutions, yellow ovals for illocutionary connections, purple ovals for discourse transitions, green ovals for inference relations, orange oval for rephrase relation, red oval for Default Conflict

the premises must be used together to support the conclusion: here, either premise, "Joan's address is 2 County Road" and "you can see a 2", cannot support "those shoes might be Joan's". However, together they do. This Default Rephrase adds additional information, although they may also reframe a proposition. In Figure 1, a proposition about Joan's shoes gives new information that the shoe prints are "probably not useful". Female 2 can be seen to agree with Female 1. Agreement and disagreement are not typically reconstructed propositionally, but the relationship is still captured through the appropriate illocutionary force. The IAT annotation guidelines are available online.[4]

3.3.2. The Data

The RIP corpus consists of three annotated subcorpora: 49,000 words total. This game was chosen due to its "Easy" difficulty setting and the amount of good reviews the product had; the game used had to be of a certain quality, couldn't be nonsensical, and had to be solvable. The conversations that the groups had were recorded, transcribed, and then annotated using Inference Anchoring Theory. All participants had never played a murder mystery game before. Each group contributed different amounts of data due to the differing amount of time Group 1, 2, and 3 spent on the game. Group 1, 9,170 words; Group 2, 13,609 words; Group 3, 26,279 words.

3.3.3. Inter-Annotator Agreement

An integral step in corpus linguistic development is to measure the reliability of the analysts in their annotation of the text. Before distribution to annotators, the transcript is segmented into parts. As is common in the use of IAT, there is a two-step method to controlling quality: during initial analysis, annotators have their own parts reviewed by another annotator, and do the same in return – this encourages discussion and catches minor mistakes; the second step is where 10% of the corpus is then re-annotated and reviewed by a different set of analysts. The original and re-annotated annotations are then automatically compared and an inter-annotator agreement (IAA) score is calculated. In the RIP corpus $\kappa = 0.68$, showing substantial agreement (Landis and Koch, 1977); therefore amongst annotators, agreement was fairly strong and the annotations of the corpus cohesive, particularly for what is a demanding and high-level pragmatic annotation task notorious for low agreement scores. This number is in line with similar corpora; Visser et al. (2022), for instance, report a kappa score of 0.61.

[4] https://www.arg.tech/index.php/annotation-guidelines/

4. Putting RIP into Perspective

4.1. Example of Data

Figure 2 shows the participants making and rejecting hypotheses. Female 1 initiated a hypothesis that Cherie did not have the opportunity to murder Nick and backed this up ("the neighbours heard Cherie"). Male 1 rejected the hypothesis as he noted that Cherie's alibi was for an hour earlier than the time of death. Female 2 continued with another disagreement aimed at Male 1. She mentioned how the evidence papers don't note if the suspect had left. Both Female 1 and Female 2's assertions become part of a linked argument, leading to the conclusion "Chris was probably in Chris' house". This is a linked argument due to how the premises work together to support Female 1's assertion. Female 1 and Female 2 both supported the hypothesis that Cherie did not have opportunity, and therefore had an alibi, through how they interpreted the evidence, whilst Male 1 rejected the hypothesis due to his own interpretation.

4.2. Failing and Succeeding

Only one group successfully completed the game. This was the group that discussed the most (26,679 words) due to the length of time it took them (two hours and forty minutes). This section will break down some of the ways the groups failed, why it's interesting, and how the final group worked their way to the correct conclusion.

4.2.1. Group 1: A Jump to Conclusions

Group 1 missed some of the evidence that gave their main suspect, Chris, an alibi. Amongst the evidence was a handwritten report from a deputy officer, who wrote their *E*'s in a particularly stylised form that made it distinct from other handwriting. This ought to have been matched to Chris' handwriting, showing that he was the deputy officer on call. Female 1 summed up their position:

(1) Female 1: *But we don't know how they did it. So, yeah, I feel like Chris is missing in action. Cherie was probably at his house, Donna was, like, had the traffic, had the police stop, Joan seems to have been at the ritual.*

Male 1 attempted to critique Cherie's alibi. Female 1 noted that Cherie had an alibi as she was making a lot of noise in a house, but Male 1 rejected this with, "But that was one hour earlier". This did not go much further, however. They discussed Cherie only a little more before moving on, and when giving their final answer – Chris and Cherie working together but Chris being the actual murderer – they did not mention at all how they discounted Cherie from this narrative. The group was on the right track: they identified Cherie as suspicious and a part of the plot, but failed to recognise her flawed alibi and that Chris had an alibi because of Donna.

4.2.2. Group 2: Failing in Two Ways

Group 2 failed in two pertinent ways: fatigue and bias. Their gameplay lasted one hour and a half, more than twice the expected time, leading to tiredness and explicitly verbalised frustration. This is a finding echoed in real investigation scenarios. Mental fatigue, whilst being difficult to quantify and a largely vague area of study (DeLuca, 2005), can cause negative effects on cognitive performance (Palmer, 2013). In a game scenario, the stakes are low. In situations where the consequence could be a wrongful arrest, this problem becomes stark.

Bias and preconceptions also played a role in the group's decision-making. Male 1 said "it should be Carmen" and Female 1 agreed by saying, "that makes a better ending". Female 1 later noted that if it was Carmen, it would be "an amazing twist". A consequence of the participants preconception of what a murder mystery is – something that shocks and surprises – is confirmation-bias. The participants realised it would be clever and the kind of reveal common to murder mysteries, and as such ignored what they first read: *there are four suspects*. When they then looked for pieces of evidence that fit this narrative, it twisted everything Carmen did as suspicious, seen in (2):

(2) Female 1: *Could we...? Before we make our declaration, can we read through Carmen's letter once more and see if we can pick anything up?*
Male 1: *Sure.*
Female 1: *Because I think she did it.*

Female 1 asked to reread a letter with the intention to find information to back up the group's hypothesis that Carmen was the murderer and suggests confirmation bias (Klayman, 1995). Approximately 9,000 words in, Male 1 explicitly mentioned that "we can't be, like, confirmation-bias-y" in regards to the group's suspicions that the murderer was Chris. The way the group failed here, reading too much into the domain-space, has its real-investigation equivalent. Statistics, such that *82% of women are murdered by men they know*[5], can both help to narrow the field of focus and create a potentially incorrect suspect pool. Interpretations of statistical evidence can be problematic, e.g., lead to the

[5] https://www.weforum.org/agenda/2020/11/violence-against-women-femicide-census/

Defence Attorney Fallacy or Prosecutor's Fallacy (Thompson and Schumann, 2017).

4.2.3. Group 3: A Lengthy Success

Group 3 is the only group that was successful in solving the game. This success is mitigated only by the length of their process: two hours and forty minutes. In real-use case scenarios, time limits are a very real restriction, whether it is because only so much man-power and time can be applied to a single case, or because there is a life or follow-up action at stake. This group had strong roles: Female 1 would run through the evidence and Female 2 would record it, whilst Female 3 tended to act independently, which included jumping in with questions as well as her own hypotheses. Female 3's question-asking tactics can be seen in Section 4.3. Female 3 also made use of a large whiteboard to keep track of means, motive, and opportunity. A timeline was also used to track each suspect's motive. Eppler and Pfister (2013) mentions police-use of "knowledge boards", which often include maps, diagrams, and sketches, etc. The group did question Carmen's influence and intentions:

(3) Female 2: *Cause if she did know, then that makes our Carmen the lead suspect now.*

However, they did not follow this lead as far as the second group did. As Female 1 says, "I don't think it's Carmen" in response to Female 3's question about who they can choose as the murderer; she was taken into consideration, but only briefly. This group was thorough in their discussion of the suspects; they often repeated found evidence and reasoning trains as Female 3 sometimes interjected with hypotheses:

(4) Female 3: *I think it's Chris who killed him. I mean, I think he's like the most straightforward and most obvious. Chris killed him and then because Cherie asked him.*

(5) Female 3: *I think Donna is in on this. I think Donna's in on this.*

The other two participants would then challenge Female 3 ("why?") and force Female 3 to explain her reasoning. This allowed the group to go through the evidence thoroughly, either to reject the hypothesis or give it grounding.

4.3. Strategies for Hypothesis-making

Group 1 and Group 2 both had similar dynamics that can be seen through the number of assertions made. Both groups had one participant who did the majority of the talking, one quiet participant who mostly contributed through agreement, and then the remaining participant who was involved but not as highly as the other participant. Group 3 breaks this pattern with similar assertion numbers (337, 476, and 333). This suggests a more even split of discussion and contribution. Their agreement level was lower than the other groups and their level of arguing was higher. The analytics for participants' illocutionary forces can be seen in Table 1.

Within group working, there is the study of group dynamics and roles (Cesareni et al., 2016; Dowell et al., 2020); whether that be on predetermined roles or those that arise naturally without prescription: 'emergent' roles. Strijbos and De Laat (2010) use four terms to describe emergent roles in group work: Captain, Free-rider, Ghost, and Over-rider. The Captain is the active and socially responsible participant; the Free-rider contributes when prompted; the Ghost barely participates; and the Over-rider attempts to realise personal goals.

An analysis of how often a suspect was mentioned, in either a premise or conclusion, was carried out in order to investigate how suspects were used within arguments. The proportions of relations including a suspect's name as antecedent or consequent out of all the relations within the different corpora is shown within Figure 3, Figure 4, and Figure 5. So, for example, propositions including Cherie's name including the use of Default Inferences are 13.95% of the entire percentage of inferences made within Group 1. Across all three groups, almost all conflicts involve reference to the suspects (96.88%, 100%, and 82.65%), whereas both inferences and rephrases made up a much smaller amount of total relations within the sub-corpora: 41.86%, 40.69%, and 35.71% for inferences, and 66.21%, 53.36%, and 52.64% for rephrases, Group 1, 2, and 3 respectively.

4.3.1. Group 1

Female 2 stood out with the highest rates of agreement: at 21.7% it was her second biggest contribution. Agreement in the whole corpus otherwise ranged from 4.4% to 14.2%. Male 3 asked a lot of questions (13.1% as opposed to 4.9% and 5.4%) and did not make a lot of arguments. He did disagree the most, at 8%. Both Female 1 and Female 2 fell beneath that number with 3.4% and 3%. Female 1 talked the most (267 assertions overall) as well as argued the most: it was her second biggest contribution at 15.7%. Female 1 could fit the role of Captain, whereas both Female 2 and Male 1 fit the Free-rider archetype; neither of the latter attempted to fulfil their own personal goals, nor were either entirely absent from the discussion.

Looking at the percentages of the relations within Group 1 as a whole, a majority of relations cover Chris or Cherie. Carmen, whom this group did not consider a suspect, is barely discussed. Both Joan

	F1.1	F2.1	M1.1	F1.2	M1.2	M2.2	F1.3	F2.3	F3.3
Asserting	50.4%	45.1%	39.2%	52.7%	46.5%	43.3%	49%	50.8%	45.2%
Arguing	15.7%	11.2%	7.4%	10.8%	10.7%	8.5%	11.2%	14.5%	11.8%
Restating	12.3%	8.1%	9.7%	11%	9.9%	8.5%	7.9%	10.9%	9.9%
Agreeing	7.9%	21.7%	13.1%	7.4%	11.6%	14.2%	10.6%	6.2%	4.4%
Pure Questioning	4.9%	5.4%	13.1%	8.5%	7%	8.5%	4.7%	6.7%	9.7%
Assertive Questioning	1.5%	1.7%	6.8%	1.8%	3.1%	1.4%	3.9%	4.1%	8.6%
Default Illocuting	3.8%	3.7%	1.7%	3.8%	5.1%	8.5%	8.1%	3.4%	3.9%
Disagreeing	3.4%	3.1%	8%	3.3%	3.7%	5%	4.1%	2.6%	4.9%
Challenging	0.2%	0%	1.1%	0.3%	2%	1.4%	0.3%	0.3%	1.6%

Table 1: Analytics of participants' illocutionary forces. Percentages are rounded to one decimal place. F and M stand for female and male with the number relative to which participant they are, and the number after the full-stop refers to which group. E.g., F1.1 is Female 1 from Group 1. Any relations below 1% are excluded from the table

and Donna are discussed very little – at 3.88% for Donna and 9.3% for Joan in regards to relations. These numbers reveal the shape of discussion within the group: they focused mainly on Chris and, to a lesser extent, Cherie, and only covered Donna and Joan to some extent. Whilst it is difficult to make generalisations, this type of discussion as well as the fact this group failed to choose the correct murderer, suggests that confirmation bias may have played a role. They settled on Chris and Cherie being the murderers and then discussed matters around the duo which built to both a stronger case, and a weaker case, as they failed to consider alternative angles.

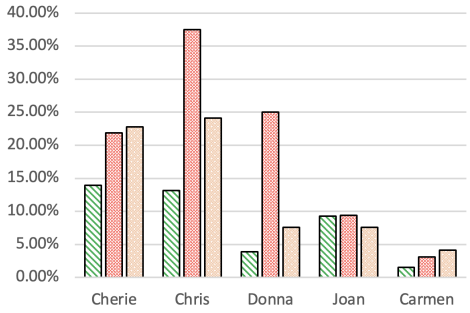

Figure 3: Relative proportions of argumentative relations spent on discussing the suspects by Group 1: first (green striped) bar showing RA proportions, second (red hatched) bar CAs, third (orange dotted) bar MAs

4.3.2. Group 2

Male 1 and Male 2's second largest contributions were agreeing (11.6% and 14.2% respectively), contrasting with Female 1's second largest contribution being restating (11%). For all of Group 2, arguing was their third largest contribution. Male 2 spoke the least, with 61 assertions in contrast to Female 1 at 321 assertions; Female 1 therefore fits the role of Captain. Male 2 asked the second most questions and answered the most questions. Male 2, spending a portion of the game attempting to crack the code puzzles, partially fits the Overrider archetype, as well as the Free-rider. Female 1 agreed the least at 7.4%.

This group and the way they failed is reflected in Figure 4. They first thought it may have been Chris who was the murderer, and this was something continued even until near the end:

(6) Female 1: *Who do you think it...? Who do you think did it? Do you not know who...? Do you not have a firm idea? The only person out of those four suspects that I think is Chris...*
Male 1: *Yes.*
Female 1: *...and if it's not him it's Carmen.*

Whilst this statement was made near the end, the group still continued with Carmen as the main suspect. This predilection for Chris as a suspect can be seen within the graphs – as well as the sudden, and late, switch to Carmen as a suspect. She did not receive a lot of attention. There were nine arguments and nine conflicts associated with her, which can be starkly contrasted to 39 rephrases associated with her; showing there were few arguments made for or against Carmen, but what was said was generalised and reframed. It also makes sense that this was a conclusion the group came to later in the game, when they were more fatigued and ready to finish the game. They settled on Car-

men despite another having suspect they strongly suspected and despite their lack of arguments for Carmen.

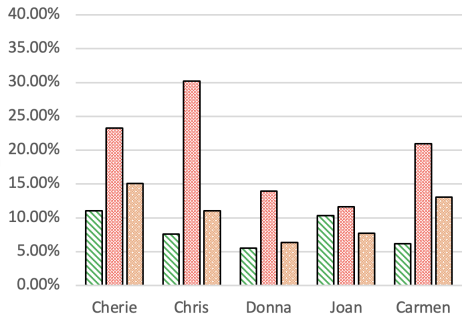

Figure 4: Relative proportions of argumentative relations spent on discussing the suspects by Group 2: first (green striped) bar showing RA proportions, second (red hatched) bar CAs, third (orange dotted) bar MAs

4.3.3. Group 3

All three participants' second largest contributions were arguing. Female 3 agreed the least, and has the lowest score within the whole corpus (4.4%). Female 1 answered the most questions within the group at 8.1%, just below her level of agreement (10.6%). This suggests she was active and socially responsible, i.e., a Captain. Female 3 asked the most Pure and Assertive Questions with 9.7% and 8.6% respectively. Participants had the lowest agreeing scores within the corpus.

Figure 5 strongly suggests that this group considered and then rejected the other suspects, before coming to the conclusion that it was Cherie. Joan and Donna, whom both Group 1 and Group 2 discussed little, featured heavily in this group's discussions. Chris was discussed similarly, which again suggests that he was considered but ultimately rejected as a suspect. Cherie (other than Carmen, whom this group considered as a suspect only briefly) received the least discussion, which suggests that this group avoided the confirmation bias that Group 1 fell into. They rejected the other suspects before choosing Cherie.

5. Relevance, Limitations, Future Work

5.1. Resource Relevance

The RIP corpus and its annotation method create a resource for training Argument Mining systems (Budzynska and Villata, 2016; Saint Dizier, 2016). Large Language Models (LLMs), while having lower requirements on data for tuning, are

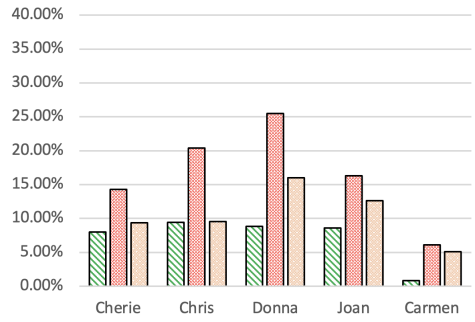

Figure 5: Relative proportions of argumentative relations spent on discussing the suspects by Group 3: first (green striped) bar showing RA proportions, second (red hatched) bar CAs, third (orange dotted) bar MAs

shown in our preliminary (yet unpublished) experiments to struggle with inference detection and do not produce state-of-the-art results. The presented corpus is aimed at providing a resource for Computational Linguistics, with a particular focus on a phenomenon and genre that are to the best of our knowledge not addressed in existing annotated corpora. The various Argument Mining tasks (Lawrence and Reed, 2020) in particular still suffer from data paucity, especially when it comes to more fine-grained annotations like the one presented here. Argument Mining systems depend upon high-quality data for proper training, which the RIP corpus provides a rich resource for.

5.2. Limitations

There are natural limitations to the domain used. A synthetic domain cannot, with total accuracy, reflect the domain it represents. A murder mystery game may take on similar attributes and use similar pieces of evidence, but it cannot accurately represent how an actual murder investigation would play out, nor all the evidence – or lack thereof – that a real team would have access to. The game must be solvable and bring an element of fun to the exercise; it is a game to be bought, played, and enjoyed. Real world events do not have neat and tidy endings that can be found with a small set of evidence and, due to the serious nature of many investigations, are lacking the element of enjoyment that a game must create.

5.3. Future Work

This corpus explores the hypothesis-making space of investigative work in a limited and finite space, capturing and discussing the small set of possible hypotheses. It puts forward a use-case in the space of collaborative working, specifically hypothesis-making. It shows promising, if preliminary, results

about the strategies of hypothesis-making which encourages future work. Future work, with restrictions in place, would be done to more strongly depict certain aspects; instead of allowing individual roles to be emergent, roles could be prescribed to the participants, or certain critical questions could be used within the course of the dialogue to create a dataset to better understand the use of questions in forming, evaluating, and rejecting hypotheses.

6. Conclusion

In recognising the lack of necessary work that combines the areas of hypothesis-making and how groups collaboratively problem solve, the RIP corpus is a novel and preliminary dataset; it captures the type of reasoning and hypothesising that happens within group work in these types of domains. It is the first time that the empirical analysis of personal analytics and groups strategies is captured within the fictitious murder investigation domain, and thus this dataset can give new insight into collaborative hypothesis generation.

This corpus reveals both the winning and losing strategies employed by the groups. Confirmation bias, the failure to consider alternative angles, and reading too much into the domain space contributed to the lack of success by two groups. There was a notably lower amount of arguing about their chosen suspect, revealed through the IAT annotation.

The group who correctly identified the culprit showed a higher proportion of arguing amongst participants, as well as higher proportions of question-asking and answering than other groups. Analysing the data also reveals how they employed a different strategy; one of discarding suspects before examining the culprit, reducing the likelihood of confirmation bias. Using empirical data to examine how participants make arguments furthers the understanding of generating hypotheses, leading to better hypothesis formation in the future.

7. Acknowledgements

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.

8. Bibliographical References

Miriam Bassok and Yaacov Trope. 1984. People's strategies for testing hypotheses about another's personality: Confirmatory or diagnostic? *Social Cognition*, 2(3):199–216.

Daniel J Benjamin. 2019. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186.

F.J. Bex. 2011. *Arguments, Stories and Criminal Evidence: A Formal Hybrid Theory*. Law and Philosophy Library. Springer Netherlands.

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. 2014. 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 and Serena Villata. 2016. Argument mining. *IEEE Intell. Informatics Bull.*, 17(1):1–6.

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.

Donatella Cesareni, Stefano Cacciamani, and Nobuko Fujita. 2016. Role taking and knowledge building in a blended university course. *International Journal of Computer-Supported Collaborative Learning*, 11(1):9–39.

John DeLuca. 2005. *Fatigue as a window to the brain*. Issues in clinical and cognitive neuropsychology. MIT, Cambridge, Mass.

Nia Dowell, Yiwen Lin, Andrew Godfrey, and Christopher Brooks. 2020. Exploring the relationship between emergent sociocognitive roles, collaborative problem- solving skills and outcomes: A group communication analysis. *Journal of Learning Analytics*, 7(1).

Martin J. Eppler and Roland Pfister. 2013. Best of both worlds: Hybrid knowledge visualization in police crime fighting and military operations.

In *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies*, page 1–8, Graz Austria. ACM.

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

Charles F. Gettys and S.D. Fisher. 1979. Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance*, 24(1):93–110.

Daniel Gigone and Reid Hastie. 1997. Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, 121(1):149–167.

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.

Markus Guhe and Alex Lascarides. 2014. Game strategies for the settlers of catan. In *2014 IEEE Conference on Computational Intelligence and Games*, page 1–8, Dortmund, Germany. IEEE.

Hayley Hung and Gokul Chittaranjan. 2010. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM international conference on Multimedia*, page 879–882, Firenze Italy. ACM.

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.

Ilias Katsagounos, Dimitrios D. Thomakos, Konstantia Litsiou, and Konstantinos Nikolopoulos. 2021. Superforecasting reality check: Evidence from a small pool of experts and expedited identification. *European Journal of Operational Research*, 289(1):107–117.

Joshua Klayman. 1995. *Varieties of Confirmation Bias*, volume 32, page 385–418. Elsevier.

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

Maria Koutsombogera and Carl Vogel. 2018. Modeling Collaborative Multimodal Behavior in Group Dialogues: The MULTISIMO Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Joshua M Lewis, Patrick Trinh, and David Kirsh. 2011. A corpus analysis of strategy video game play in starcraft: Brood war. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33.

Laura K. Palmer. 2013. The relationship between stress, fatigue, and cognitive functioning. *College Student Journal*, 47(2):312 – 325.

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.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn discourse treebank version 3.0.

Chris Reed and Katarzyna Budzynska. 2011. How dialogues create arguments. In *Proceedings of the 7th conference on argumentation of the International Society for the Study of Argumentation*, pages 1633–1645.

Jeremy Roschelle and Stephanie D. Teasley. 1995. *The Construction of Shared Knowledge in Collaborative Problem Solving*, page 69–97. Springer Berlin Heidelberg, Berlin, Heidelberg.

Patrick Saint Dizier. 2016. Argument mining: the bottleneck of knowledge and language resources. In *10th International conference on language resources and evaluation (LREC 2016)*, pages pp–983.

Noor Shaker, Stylianos Asteriadis, Georgios N. Yannakakis, and Kostas Karpouzis. 2011. *A Game-Based Corpus for Analysing the Interplay between Game Context and Player Experience*, volume 6975 of *Lecture Notes in Computer Science*, page 547–556. Springer Berlin Heidelberg, Berlin, Heidelberg.

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*.

Robert Slavin. 2011. Instruction based on cooperative learning. *Handbook of Research on Learning and Instruction*.

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.

Jan-Willem Strijbos and Maarten F. De Laat. 2010. Developing the role concept for computer-supported collaborative learning: An explorative synthesis. *Computers in Human Behavior*, 26(4):495–505.

P.E. Tetlock and D. Gardner. 2016. *Superforecasting: The Art and Science of Prediction*. Random House Business.

William C. Thompson and Edward L. Schumann. 2017. *Interpretation of Statistical Evidence in Criminal Trials*, 1 edition, page 371–391. Routledge.

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2022. *Annotating Argument Schemes*, page 101–139. Springer Nature Switzerland, Cham.

Simon Wells and Chris A Reed. 2012. A domain specific language for describing diverse systems of dialogue. *Journal of Applied Logic*, 10(4):309–329.

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