Identifying Questions Under Discussion in Naturalistic Discourse

Karl Mulligan and **Kyle Rawlins**
Department of Cognitive Science
Johns Hopkins University
karl.mulligan@jhu.edu

1 Introduction

The Question Under Discussion (QUD) model of discourse structure has been an influential theoretical device in the formal pragmatics toolkit, but efforts to derive QUDs from naturalistic data are few. In this work, we crowdsource QUD annotations of radio interviews with many ($N = 10$) annotators per sentence. We explore a fundamental issue underlying QUD theory: can discourse agents reliably infer implicit questions being addressed in naturalistic discourse? We also investigate whether, as most QUD theories presuppose, there is at most a single immediate (i.e., current, top-of-stack) QUD at a given turn of discourse. We compare several similarity metrics for questions and answers, demonstrating that our user interface encourages annotators to obey useful theoretical constraints like Question-Answer Congruence (Riester, 2019). Overall, we find moderate annotator agreement forming qualitatively identifiable clusters, consistent with the existence of multiple contextually-restricted immediate QUDs. We further find, unexpectedly, that annotators are unreliable at reconstructing masked overt questions, suggesting that explicitly asked questions may correspond to shifts in QUD or topic.

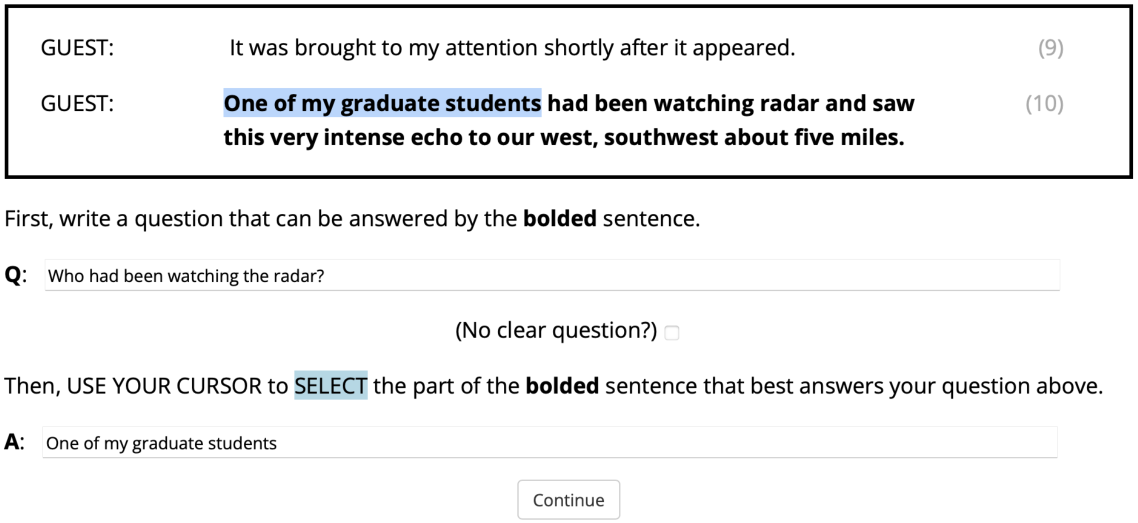
2 Background

The idea that much of discourse can be structured by implicit questions is part of many theories of pragmatics (Van Kuppevelt, 1995; Ginzburg, 1996), but it is perhaps most clearly and influentially articulated in the framework of Roberts (2012/1996). Roberts characterizes discourse as a game in which possible moves (utterances) are guided by whether they help answer the immediate QUD, usually a single implicit question assumed to be shared by discourse participants. QUDs are thus a useful way of operationalizing the Maxim of RELATION (Grice, 1975): an utterance is relevant in a discourse context if and only if it addresses the immediate QUD. QUDs play a central role in formal analyses of a wide variety of pragmatic phenomena, including focus and contrastive topic (Roberts, 2012/1996; Büring, 2003), not-at-issue content (Simons et al., 2010), implicature (Van Kuppevelt, 1996; Degen, 2013), particles like German *doch* (Rojas-Esponda, 2014), exclusives like *only* (Coppock and Beaver, 2014) and *just* (Warstadt, 2020), overall discourse structure (Ginzburg, 1996), and many more. Yet the QUDs used in motivating these accounts are often generated *ad hoc*, taking for granted many assumptions about about the nature and availability of QUDs in discourse. As such, to investigate these assumptions, we direct our efforts toward the task of collecting QUDs dervied from naturalistic dialogue, from theoretically naive annotators.

2.1 Prior work

Existing annotated resources fall broadly into two camps: rigorous, theory-grounded annotation approaches, such as the hierarchical annotations in De Kuthy et al. (2018) and Hesse et al. (2020), albeit limited in scope by ontological complexity; or large, crowdsourcing approaches working with various kinds of implicit question, such as evoked questions (Westera et al., 2020) or elaboration questions (Wu et al., 2023), albeit not necessarily target-

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 357-361.
Irvine, California, June 27-29, 2024



GUEST: It was brought to my attention shortly after it appeared. (9)

GUEST: **One of my graduate students had been watching radar and saw this very intense echo to our west, southwest about five miles.** (10)

First, write a question that can be answered by the **bolded** sentence.

Q: Who had been watching the radar?

(No clear question?) □

Then, USE YOUR CURSOR to SELECT the part of the **bolded** sentence that best answers your question above.

A: One of my graduate students

Figure 1: Annotation interface. The answer box is auto-populated only by selecting from the bolded (target) sentence.

ing theoretical properties of immediate QUDs. Most recently and closest to this work, Wu et al. (2024) collect data about the salience of implicit questions in a given context, but focusing on inquisitive questions rather than immediate QUDs, which serve related but distinct explanatory purposes for discourse. Furthermore, these latter resources are predominantly sourced from written texts or monologues, which may not be fully representative of the kinds of QUDs and QUD transitions that occur in naturalistic two-party dialogue.

3 Methods

3.1 Procedure

We selected 10 complete two-party dialogue transcripts from INTERVIEW (Majumder et al., 2020), a corpus of National Public Radio (NPR) interviews in American English, split by sentence and annotated with turn information. Episodes were chosen to have between 29 and 32 sentences, of which at least 5 were overt questions ($\mu = 5.5$). In order to get a sense of QUD variability, we showed each interview to a high number of annotators: 10 native English speakers per episode were recruited on Prolific, resulting in a total of 100 unique sets of annotations. For each episode, annotators read the dialogue one sentence at a time, in a moving two-sentence window to simulate linear processing, as inspired by Westera et al. (2020). For each new sentence, annotators were prompted to (i) write a question that can be answered by that sentence, and (ii) select a contiguous span from that sentence best representing the answer to their question (Figure 1). Annotators could opt to mark "no clear question" (e.g., for non-declarative moves like *Good morning.*) While participants were free to write any question that the sentence addresses, we assume that discourse context makes certain potential QUDs more likely. (Indeed, we find QUD variability is not modulated by sentence length.)

3.2 Evaluation

We consider several similarity metrics for measuring QUD agreement. The first is *token edit distance* (ED), which counts the minimum number of words (tokens) that must be inserted, deleted, or substituted to transform one array of tokens into another. This metric is useful for measuring answer similarity ($\mu = 6.6$), since all answers are forced by our interface to be subsets of the target sentence. Directly measuring similarity among questions is more challenging, due to the open-ended nature of the prompt. By assuming Q-A Congruence, we hypothesize that annotators who select similar (low ED) answer spans are more likely to be writing similar QUDs, since they place

Metric	((1),(2))	((2),(3))	((1),(3))
Token ED (A)	3	4	5
Token ED (Q)	6	8	7
BERTScore (Q)	0.41	0.14	0.12
Wh-word (Q)	1	0	0

Table 1: Similarity metrics for answer spans (A) and questions (Q) on annotator-written QUDs (1) – (3).

focus on the same information. As such, we expect a good question similarity metric to be correlated with our answer span similarity metric. To test this hypothesis, we look at how answer ED correlates with three question similarity metrics. The first is *question edit distance* ($\mu = 8.0$), which is simply token edit distance applied to questions. The second is rescaled *BERTScore* (Zhang et al., 2020) ($\mu = 0.37$), which encodes two sentences using a large transformer language model and measures the cosine similarity of their embeddings (between –1 and 1, where higher values are more similar); neural similarity metrics are more robust to disparities in surface form and are useful for capturing intuitive notions of similarity like synonymy. The third metric is *Wh-word agreement*, which is a boolean measure returning true when both questions have the same Wh-word (or auxiliary, for polar sentences) ($\mu = 0.39$). Examples of these metrics applied to pairs of the collected data below in (1) – (3) are given in Table 1 (see Figure 1 for the source utterance).

(1) Who else had been watching the radar? [**One of my graduate students**]

(2) Who saw the occurrence and effects on the radar? [**my graduate student**]

(3) Where are the clouds coming from? [**southwest about five miles**]

With these metrics, we intend to capture the intuition that (1) and (2) should be considered essentially the same QUD (i.e., targeting the same information structure), whereas (3) is asking something quite different from each.

4 Results

We find a moderate correlation for answer ED and question ED (Spearman's $\rho = 0.41$), as well as for answer ED and BERTScore using DeBERTa ($\rho = -0.37$), the model recommended by the BERTScore authors. These correlations suggest that the data produced by the annotators obey Q-A Congruence, an important property of immediate QUDs which connect questions to the parts of the utterance which answer them.

We also find correlations for Wh-word agreement with answer ED ($\rho = -0.32$) and BERTScore ($\rho = 0.49$), which taken together highlight the importance of the Wh-word as a signal for the selection of ansewr spans and overall question similarity. The vast majority of QUDs written are Wh-questions, though the polar questions produced exhibit an interesting pattern. With no explicit instruction to do so, for polar QUDs, annotators often select the entire sentence as their answer span (a response consistent with theoretical predictions about focus), while Wh-QUDs have short, constituent-sized spans.

4.1 Masking questions

Under most theories of QUDs, in normal circumstances, explicitly asked questions become the new QUD. To see whether annotator-written QUDs match actual questions from the interview, we masked each explicitly asked

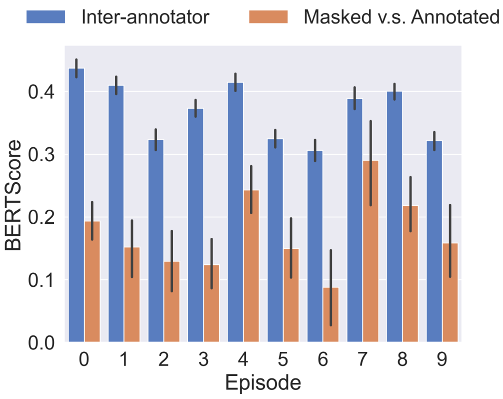


Figure 2: Mean question similarity across annotators (blue) and between the masked question and annotator-written questions (orange).

question by replacing it with "[QUESTION MASKED]" while keeping the sentence preceding it intact for context. We found that across episodes, annotators write QUDs consistently less similar to the masked question than to one another (Figure 2), yet the mean inter-annotator BERTScore for QUDs on post-masked trials is not significantly different from inter-annotator agreement on normal trials.

5 Discussion

Overall, we find that crowdsourced annotators are capable of producing immediate Questions Under Discussion, exhibiting moderate agreement, but still with quite variable annotations. What accounts for this variability? One explanation may be a design concern, namely the lack of specificity in the prompt: we ask for *a* question that can be answered by the target sentence, rather than directly requesting "the most relevant" or "the most important" question. But if agents do (at least implicitly) track QUD as part of a model of discourse, we would assume that the "true" QUD is the most salient and cognitively accessible question.

Alternatively, the variability may stem from theoretical concerns, i.e., assumptions about QUD theory which must be relaxed to account for discourse in practice. For instance, the model of Roberts (2012/1996) assumes that there may be at most one immediate QUD at any given turn in discourse. However, one possibility consistent with our data is that there is *inherent multiplicity* of QUDs, targeting multiple parts of the utterance at once: some information may be privileged over others, but not exclusively so. This may especially be the case for complex, multi-clause utterances. Another possibility is that the variability reflects agent *uncertainty* about the QUD; discourse context may not always sufficiently constrain the QUD, an idea which is related to Stalnaker's "defective context" or uncertainty about the state of the conversational scoreboard in the sense of Lewis (1979). Our methods in this work are limited in their ability to tease apart inherent multiplicity from uncertainty, but we believe this distinction is an important one to make in future work.

As for the recovery of masked questions, the data appear to challenge the assumption, made by many theories of discourse, that explicitly asked questions become by default the immediate QUD in a way that is fully recoverable by a hearer. One explanation is that discourse participants may opt to ask explicit questions precisely in contexts with unpredictable topic shifts, making recovery difficult. Another is that our question similarity metrics fail to distinguish between the immediate QUD and potentially more general superquestions, a limitation of our present focus on immediate QUDs rather than hierarchical representations of discourse structure. This problem may be amplified by genre effects, since responses to explicit questions in NPR interviews tend to involve multi-sentence turns often beginning with some kind of exposition; the immediate QUD of the initial utterance may therefore be a subquestion of the the asked question, but be unfairly counted as distinct by our metrics. As such, we are unable to rule out the question-to-QUD assumption outright, at least not without a more refined metric capable of detecting subquestion relations.

6 Conclusion

Our results suggest that naturalistic discourse involves multiple compatible QUDs, but annotators are able to extract these immediate QUDs. Our similarity metrics for questions and answer spans indicate that annotators produce QUDs which obey Q-A congruence, making our data potentially useful to researchers interested in focus, alternative semantics, and other QUD-sensitive phenomena. Fully characterizing the space of possible QUDs, however is limited by existing metrics and non-hierarchical representations. Implementing more nuanced question relations and question clustering methods are challenges we leave to future work.

References

Daniel Büring. 2003. On D-Trees, Beans, And B-Accents. *Linguistics and Philosophy*, 26(5):511–

545.

Elizabeth Coppock and David I. Beaver. 2014. Principles of the Exclusive Muddle. *Journal of Semantics*, 31(3):371–432.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. QUD-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Judith Degen. 2013. *Alternatives in Pragmatic Reasoning*. Ph.D. thesis, University of Rochester.

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. In Jerry Seligman, Dag Westerståhl, and Lawrence Cavedon, editors, *Logic, Language, and Computation*, volume 1 of *CSLI Lecture Notes*, pages 221–237. CSLI Publications, Stanford, California.

Herbert P Grice. 1975. Logic and conversation. In *Speech Acts*, pages 41–58. Brill.

Christoph Hesse, Anton Benz, Maurice Langner, Felix Theodor, and Ralf Klabunde. 2020. Annotating QUDs for generating pragmatically rich texts. In *Proceedings of the Workshop on Discourse Theories for Text Planning*, pages 10–16, Dublin, Ireland. Association for Computational Linguistics.

David Lewis. 1979. Scorekeeping in a Language Game. In Rainer Bäuerle, Urs Egli, and Arnim von Stechow, editors, *Semantics from Different Points of View*, Springer Series in Language and Communication, pages 172–187. Springer, Berlin, Heidelberg.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding. In *EMNLP 2020*, pages 8129–8141.

Arndt Riester. 2019. Constructing QUD Trees. In Malte Zimmermann, Klaus Von Heusinger, and Edgar Onea Gaspar, editors, *Questions in Discourse: Volume 2: Pragmatics*, volume 36 of *Current Research in the Semantics / Pragmatics Interface*, pages 164–193. Brill.

Craige Roberts. 2012/1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. 5.

Tania Rojas-Esponda. 2014. A QUD account of German 'doch'. *Proceedings of Sinn und Bedeutung*, 18:359–376.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. *Semantics and Linguistic Theory*, pages 309–327.

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

Jan Van Kuppevelt. 1995. Discourse Structure, Topicality and Questioning. *Journal of Linguistics*, 31(1):109–147.

Jan Van Kuppevelt. 1996. Inferring from Topics: Scalar Implicatures as Topic-Dependent Inferences. *Linguistics and Philosophy*, 19(4):393–443.

Alex Warstadt. 2020. "Just" don't ask: Exclusives and potential questions. *Proceedings of Sinn und Bedeutung*, 24(2):373–390.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1118–1127, Marseille, France. European Language Resources Association.

Yating Wu, Ritika Mangla, Alexandros G. Dimakis, Greg Durrett, and Junyi Jessy Li. 2024. Which questions should I answer? Salience Prediction of Inquisitive Questions.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT.