Analyzing naturally-sourced Questions Under Discussion

Karl Mulligan & Kyle Rawlins[*]

Abstract. The Question Under Discussion (QUD) framework of discourse has been a highly influential theoretical device in many accounts of various pragmatic phenomena, yet there has been comparatively little work assessing the extent to which the QUD can be reliably inferred from naturalistic contexts. In this paper, we focus primarily on measuring the *variability* across individuals in QUD inference, while also verifying other related, commonly held assumptions about QUD theory. To this end, we collect QUDs from many theoretically naive subjects tasked with processing a radio interview utterance by utterance. We consider various analyses designed to address the problem of measuring question similarity. Overall, we find that there exists moderate variability among subjects, consistent with possibly the insufficiency of context in determining QUD, or possibly also the simultaneous coexistence of multiple valid QUDs. To more adequately tease apart these possibilities, we also propose additional analyses for addressing the issue of question identity.

Keywords. Question Under Discussion; pragmatics; discourse structure; question similarity

1. Introduction. One of the most influential paradigms in pragmatics is the Question Under Discussion framework, which analyzes discourse as a process guided by addressing implicit questions (Van Kuppevelt 1995, Roberts 1996/2012). The distinguishing feature of this framework is that contextual relevance is modeled using natural language questions. One advantage of this is that the framework is able to profit from an existing rich and successful tradition in formal semantics of analyzing questions as sets of alternatives (Hamblin 1973). But another advantage of using questions is the fact that natural language questions are, of course, easily understood and generated by non-specialists. We exploit this latter advantage to investigate our central question: given the same access to discourse context, do normal language users reliably infer the same implicit QUD?

Underlying most formal accounts of QUD-sensitive phenomena is the assumption that the immediate QUD is accessible, or at least inferrable, for all discourse participants at each stage of conversation. While this may be a reasonable assumption, there is limited work so far exploring the extent to which this underlying assumption holds in natural discourse. We also have only limited evidence that the kinds of QUDs posited to account for various pragmatic inferences are indeed observed "in the wild" within typical conversation. We are thus motivated to collect a variety of QUDs from naturalistic discourse, using natural language questions produced by theoretically naive subjects. Given a sense of what such naturally-sourced QUDs are like, this may help determine the ecological validity and generality of accounts which depend on QUDs.

In this work, we break down the fundamental issue of QUD accessibility into three more specific assumptions about QUDs commonly found in the literature:

[*] We would like to thank our anonymous reviewers, as well as audiences at MACSIM, ELM, and SCiL for their feedback. Authors: Karl Mulligan, Johns Hopkins University (karl.mulligan@jhu.edu) & Kyle Rawlins, Johns Hopkins University (kgr@jhu.edu).

- SINGULARITY: There is a single immediate QUD at any given turn in discourse.
- Q-TO-QUD: An explicitly asked question becomes the new immediate QUD.
- Q-A CONGRUENCE: The answer to the immediate QUD is contained within the asserted utterance, with the Wh-word *congruent*, or corresponding, to the focused constituent (Rooth 1992).

To test these assumptions, we assess to what extent ordinary language users are able to produce and agree upon the immediate QUD while processing naturalistic dialogue. Building on data collected by Mulligan & Rawlins (2024), we find that there is persistent variability across participants in their inferred QUDs, though the data are also consistent with frequent, non-random pockets of agreement among participants, a finding which calls SINGULARITY into question but stops short of disproving it. We also find results suggesting that Q-TO-QUD may be too strong of an assumption. For the last assumption of Q-A CONGRUENCE, however, we find that our subject-generated QUDs are largely congruent to selected answer spans, which are usually grammatical constituents. Overall, we find that some contextually-constrained notion of QUD is accessible and shared across language users, but we are limited by our methods and analyses in our ability to narrow down the source of this variability.

2. Background. Although there are various theories of discourse that depend heavily on some notion of implicit question (Van Kuppevelt 1995, Ginzburg 1996), this work primarily focuses on the model described in Roberts (1996/2012). For Roberts, discourse is likened to a game, in which the various moves (utterances) are recorded on a public scoreboard. In addition to the Common Ground, and the sets of explicitly uttered assertions and questions, there is another, generally implicit component, the Question Under Discussion stack. A move in discourse is deemed relevant if it addresses the topmost item on this stack, the immediate QUD.

In principle, the next QUD can be *inferred* as a function of interlocutor goals and the information in this discourse data structure: prior conversational moves, shared Common Ground assumptions, and existing QUDs on the stack (Cooper et al. 2000, Velleman & Beaver 2016). However, the exact process by which this inference is assumed to take place is far from fully understood. QUD annotations are thus a potentially valuable resource for this effort.

There exist several works, from both linguistics and natural language processing, that attempt to source implicit questions from natural language. The approaches vary along several dimensions. Some works, like De Kuthy et al. (2018), involve detailed, hierarchical annotation paradigms that hew closely to the formalism described by Roberts (1996/2012), thus requiring trained annotators familiar with QUD theory. Other works use crowdsourced participants (Westera et al. 2020, Pyatkin et al. 2020, Wu et al. 2023), though often using more general notions of implicit question agnostic to specific structural assumptions.

Common to all of the approaches mentioned above is the fact that the source material is monologue, and typically written rather than spoken. Moreover, prior efforts in this area are generally limited to one or two annotators per item, making quantifying variability difficult, though De Kuthy et al. (2018) among others report frequent qualitative discrepancies in annotation decisions.

3. Methods. In this work, we sourced our QUDs from dialogue, and in order to quantify variability in QUD inference, we crowdsourced a large number of implicit questions for each utterance.

3.1. MATERIALS. For our conversation data, we picked 10 two-party interviews from the INTERVIEW dataset (Majumder et al. 2020), a collection of National Public Radio interviews in American English. While perhaps not as spontaneous as some other genres of speech, these interviews contain few disfluencies and are of consistently high quality, facilitating sentence-by-sentence annotation. Interviews were also generally conversational in tone, making them suitable for studying the processing of naturalistic discourse. For consistency, we chose interviews with between 29 and 32 sentences, at least 5 of which were explicitly asked questions.

3.2. PARTICIPANTS. For each interview, we recruited 10 native English speakers per episode via Profilic, for a total of 100 unique sets of QUD annotations. This high number of participants per episode allows us to get a highly calibrated measure of variability in QUD selection across participants at the utterance level. Participants took approximately 20 minutes to complete each interview, and were compensated $15 per hour for their work.

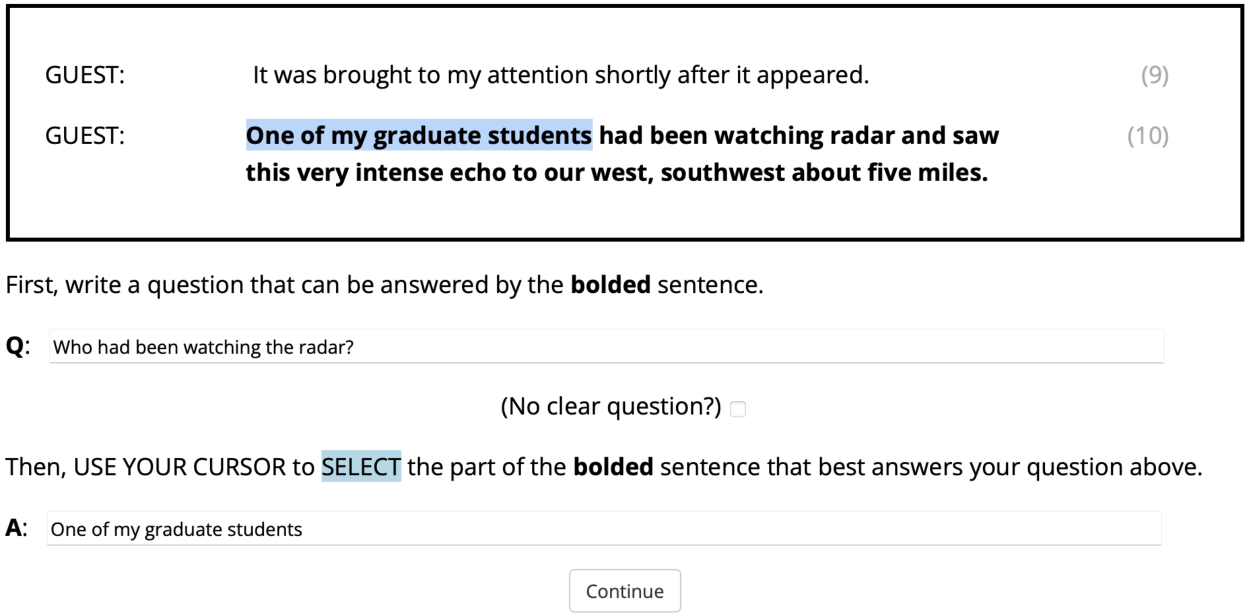
3.3. PROCEDURE. Participants were presented the interview in a moving two-sentence window in order to simulate real-time, incremental revelation of context, following Westera et al. (2020). The first of the two sentences was the *context* sentence, while the second was the *target* sentence, always indicated in bold. For each sentence pair, they were instructed to do two things: first, to write a question that could be answered by the target sentence; and second, to select the shortest contiguous span of words from the target sentence which best answers the question they just wrote. For utterances which did not serve as answers to a clear question (such as commands, requests, greetings, and other non-declarative sentences), participants were instructed to check a box labeled "No clear question" and, instead of giving a question–answer pair, to explain why there was no clear question being addressed.

As a complimentary task testing the Q-TO-QUD assumption, we also masked explicitly asked questions to see whether subjects would produce QUDs for the subsequent utterance resembling the asked question. For each literal question asked in the interview, we replaced it with "[QUESTION MASKED]" while keeping the utterance before it intact and visible as the context sentence. These trials were not any different otherwise from regular trials.

3.4. EVALUATION. In order to measure the variability in inferred QUDs across subjects, we employed three question similarity metrics, each with its own advantages and disadvantages.

The first of these is *word edit distance*, which is a sum of the minimum number of edits (insertions, deletions, or substitutions of words) needed to transform one sentence into another. Under the assumption that similar questions can be addressed by similar answers, we we also use edit distance to measure similarity between answer spans. This makes answer edit distance a useful proxy for question similarity (indeed, we find that these measures are correlated), and also more reliable since all answers for a given trial are subsets of the same utterance.

The second metric is *BERTScore*, a general-purpose neural sentence similarity metric (Zhang et al. 2020). BERTScore uses a large transformer language model to encode each question into a high-dimensional vector space, and then measures the cosine similarity of these encodings; higher values indicate greater semantic similarity. Unlike edit distance, neural metrics like BERTScore are less sensitive to variation in the surface input, and are thus more suitable for capturing paraphrases. Here we used a rescaled version of DeBERTa, the default configuration suggested by the



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: Response interface from Mulligan & Rawlins (2024). The question box (**Q**:) is a free response prompt. The answer box (**A**:) can only be filled by selecting from the target sentence, ensuring that a subject's written QUD is addressable using a contiguous span of the target utterance.

BERTScore authors.

The third metric is *Wh-word agreement* which we define as 1 if both questions share the same Wh-word (or in the case of polar questions, the same first word auxiliary), and 0 otherwise. While simple, this metric is useful for coarsely differentiating questions seeking distinct types of information.

To illustrate the use of these metrics on real data, we apply each metric to pairs of the following collected QUDs for the utterance in Figure 1, repeated from Mulligan & Rawlins (2024). The values for each metric are displayed in Table 1.

(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**]

Mulligan & Rawlins (2024) find that all of these metrics are moderately correlated with one another. The results going forward will mostly use BERTScore to measure QUD variability because of its granularity and robustness to synonymy.

4. Results.

4.1. QUD VARIABILITY. We turn now to an assessment of the first of our three assumptions, the SINGULARITY of QUD at a given stage of discourse. If there is at most a single QUD at any given time, we would expect less pairwise variability among subjects, since the inference of the current

Metric	((1),(2))	((2),(3))	((1),(3))
Word edit distance (A)	3	4	5
Word edit distance (Q)	6	8	7
BERTScore (Q)	0.41	0.14	0.12
Wh-word agreement (Q)	1	0	0

Table 1: A comparison of similarity metrics for both answer spans (A) and questions (Q) on collected QUDs. In this example, we wish to capture the qualitative intuition that (1) and (2) are most similar to each other, and are thus candidates for being qualitatively characterized as "the same" QUD.

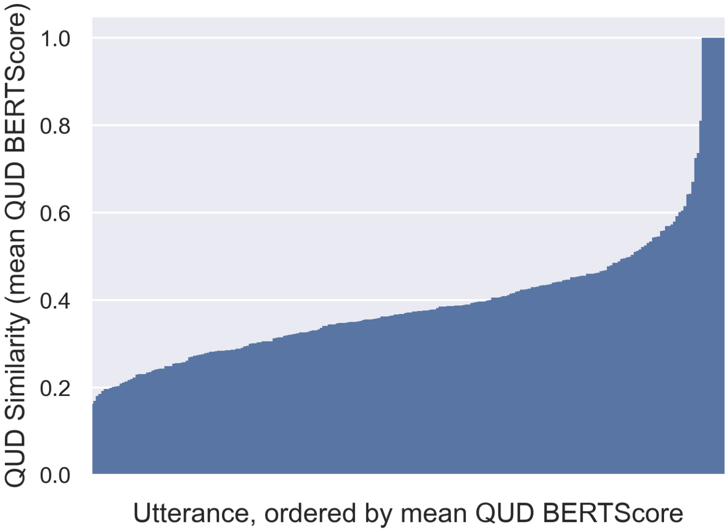


Figure 2: Each bar represents the average pairwise variability among all ten subjects for a single utterance ($N = 241$ observations), with items ordered by the mean BERTScore.

QUD should be transparent from discourse context.

We find that there is moderate variability among subjects in QUD inference, calling a strict notion of SINGULARITY into question. To visualize this overall variability, we plot the distribution of the mean BERTScore per item, ordered by ascending value, in Figure 2. A concave distribution for this plot would indicate high overall support for SINGULARITY, since the QUDs for the majority of utterances would have higher mean pairwise similarity, and the distribution would have more overall weight. However, what we in fact see is a somewhat convex distribution: the majority of trials have a mean pairwise BERTScore around 0.3, indicating only moderate similarity across QUDs.

For a qualitative view of this variability, as well as more examples of interpreting BERTScore values, we present in the Appendix a sample of QUDs from a low-agreement trial (Table 2, mean BERTScore of 0.23) as well as a sample from a high-agreement trial (Table 3, mean BERTScore of 0.49).

What is the source of this variability? This is a difficult question to answer given our data, but there are at least a few possible explanations that can be ruled out. For instance, QUD variability does not seem to be correlated with time (i.e., position in the interview). One might suppose that overall QUD variability decreases over the course of a dialogue, as discourse context accumulates and constrains the space of relevant questions; alternatively, one might instead expect overall QUD variability to increase, as more content and discourse referents are introduced. In fact, neither trend is the case: QUD variability does not trend significantly in one direction or the other over time.

4.2. MASKED QUESTION ANALYSIS. In order to assess our second commonly encountered assumption, Q-TO-QUD, we compare masked explicitly asked questions to subject-written QUDs for the utterances following them. Q-TO-QUD says that, under typical circumstances (i.e., for accepted, canonical questions) an explicitly asked question becomes the new QUD. We find that across the board, for post-masked question trials, subject QUDs are consistently less similar to the masked question (mean BERTScore of 0.172) than they are to one another (mean BERTScore of 0.369).

This finding seems to suggest that Q-TO-QUD should not be taken for granted as a default assumption, albeit with some caveats. There may be unaccounted for properties of explicitly asked questions which make them harder to reconstruct than ordinary QUDs; for instance, interlocutors may be more likely to ask an explicit question during a topic shift or when seeking information about an as-yet unmentioned entity. It may also be the case that our setup encourages subjects to write questions which reuse words from the target utterance, thereby artificially positively biasing inter-subject comparisons over masked question comparisons.

4.3. CONSTITUENCY ANALYSIS. For the assessment of our third and final assumption, Q-A CONGRUENCE, we perform a constituency analysis of the selected answer spans to see whether they are syntactically congruent to subject-written QUDs.

A possible concern over sourcing QUDs from non-specialists, as we have here, is that the resulting questions and their answer spans may not reflect essential formal properties of QUDs, such as the connection to focus and designated alternative sets. To determine whether our sourced QUDs obey the property of Q-A CONGRUENCE, we perform a constituency analysis on all utterances and answer spans to see whether subjects generally pick grammatical constituents as answers. Upon parsing these structures, we can then check whether the syntactic categories of constituents correspond to the appropriate Wh-word.

To parse the utterances, we used the `constituent-treelib` library, a neural constituent parser based on the Berkeley Neural Parser and spaCy (Halvani 2024). For each utterance in each interview, we perform a constituent parse and recursively obtain a set of all constituents in the sentence labeled by syntactic category.

We find that just under half of the answer spans are constituents, when directly compared with the outputs of our constituency parser. This is less than expected if we would like to claim that Q-A CONGRUENCE holds in our data. However, due to limitations of using an automatic parser, we notice that there are many false negatives. For instance, an answer span might not be counted as a constituent if the span fails to include an optional adjunct clause found in the original sentence, even if it is perfectly capable of standing alone as a syntactically complete noun phrase. To remedy this problem, we define the notion of *constituent leaf coverage* for a given answer span, sentence, and constituent parse:

$$\text{Constituent Leaf Coverage} = \frac{\text{\# of answer tokens}}{\text{\# of leaves in smallest common parent subtree}}$$

This notion gives us a more flexible, gradient measure of constituency: spans which are "near-constituents" are defined as those which are syntactically close to a phrasal constituent (i.e., they are contained by a common parent subtree of nearly the same size). For these answer spans, the ratio will be close to 1; for answer spans which perfectly match a constituent, this measure will be exactly 1. On the other hand, for answer spans which cross large syntactic boundaries, this ratio will be lower. In Figure 3, we see that while half of answer spans are constituents, the distribution of leaf coverage values for non-constituents skews toward 1, indicating that even non-constituent spans are likely to be syntactically informative.

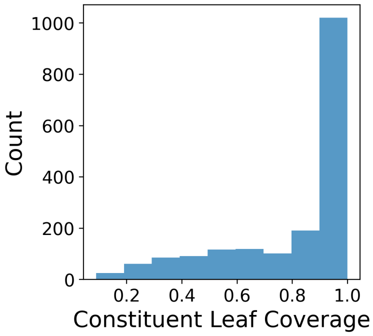
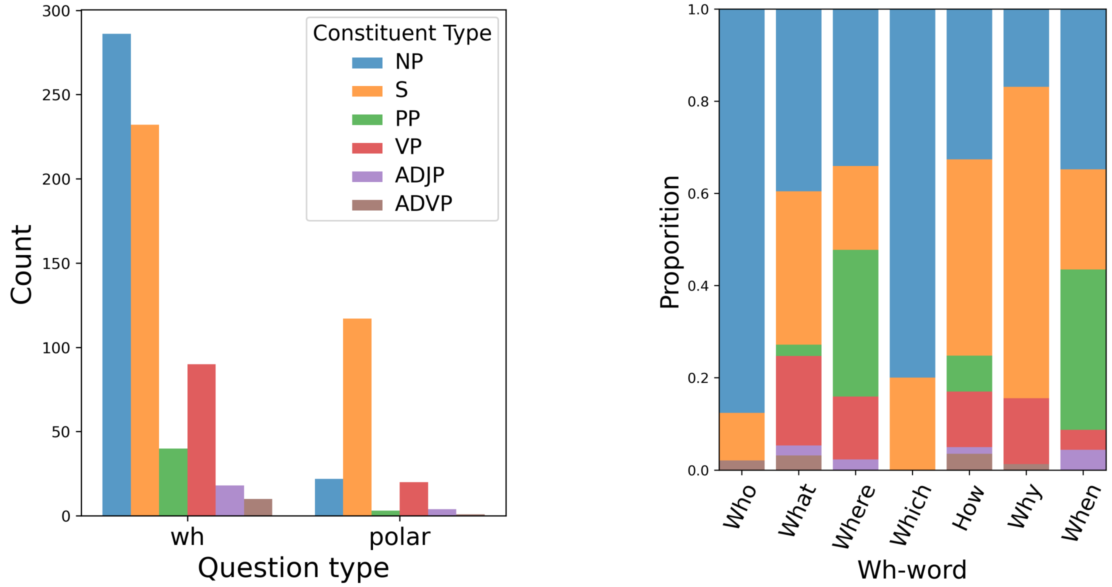


Figure 3: Distribution of constituent leaf coverage for all answer spans.

To see whether Q-A CONGRUENCE holds in our data, we examine which constituent phrase types pattern with which kinds of Wh-words (or auxiliaries, for polar questions). For these analyses, we limit our answer spans to only perfect constituents. About 70 percent of our QUDs are Wh-questions, while 21 percent are polar questions (we set aside the remaining 9 percent of questions which contain neither a Wh-word nor an auxiliary as their first word). In Figure 4a, we show a breakdown of answer span syntactic phrasal categories, separated by QUD question type. Polar questions are shown to be most often answered by entire sentences, while Wh-questions are answered by a greater variety of phrase types depending on the particular Wh-word. Figure 4b reveals how the distribution of phrase type varies by Wh-word. We see that QUDs containing Wh-words which mostly pick out sets of entities, such as *who*, *what*, and *which*, are largely composed of noun phrases; words like *where* and *when* contain a noticeably higher relative share of prepositional phrases; and words like *how* and *why* are mostly answered using full sentences. Taken together, the results of this analysis indicate support Q-A CONGRUENCE in our sourced QUDs.

5. Discussion. Our results show that, across participants, there exists moderate variability in inferring the current QUD in naturalistic discourse. Based on our large sample of QUDs per utterance, we are able to find qualitative pockets of agreement, with our question similarity metrics reflecting levels of agreement above random, though still far from universal agreement. From these analyses alone we are limited in our ability to determine the source of the variability. Still, it is worth discussing the possible causes of this variability and how these causes might be uncovered in future work.



(a)

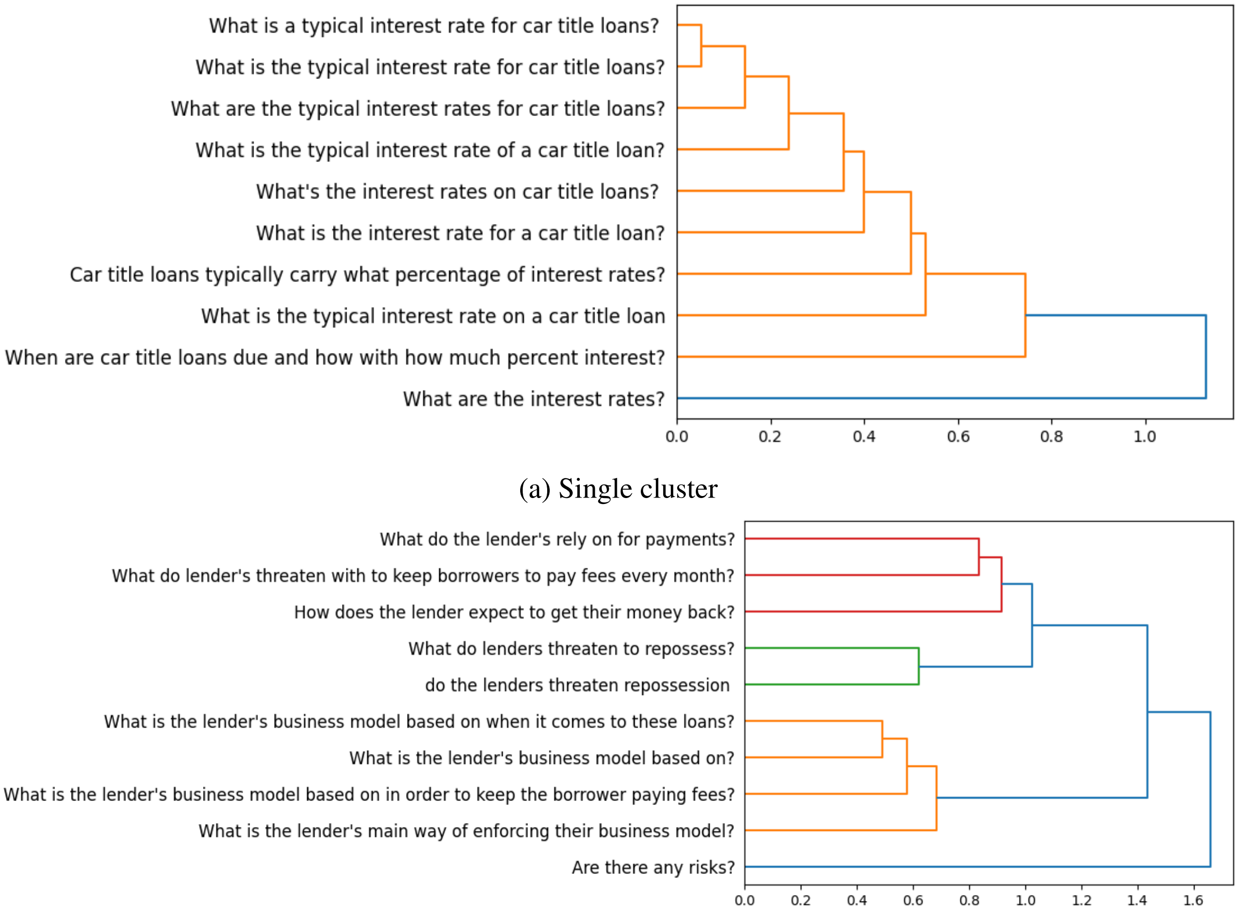
(b)

Figure 4: (a) Distribution of constituent types for Wh-questions and polar questions. (b) Proportion of answer span constituent types by Wh-word.

The data is consistent with at least two general possibilities: there may be *uncertainty* about the current QUD (assuming there is only one intended or "true" QUD) and discourse context is simply insufficient in narrowing it down; or there may be an inherent *multiplicity* of discourse-relevant QUDs, in which case SINGULARITY is too strict of an assumption and must be relaxed. Of course, these possibilities are not mutually exclusive. There may be multiple active QUDs, but with varying degrees of discourse relevance and accessibility; the perceived levels of activation for these QUDs may vary based on individual differences and discourse goals. Indeed, this sort of multiplicity is consistent with a view of QUD tracking in which questions pushed onto the stack earlier in discourse are still able to be directly addressed, something which can be captured using a hierarchical model.

To address the issue of whether there are multiple simultaneous QUDs, *clustering* may be a useful, unsupervised method of analysis. Agglomerative (or bottom-up) hierarchical clustering starts with each data point as its own cluster, and at each step, the two closest (most similar) clusters join, resulting in a a hierarchically-ordered grouping of similar data points.

While we do not perform any extensive clustering analysis in this paper, we would like to share some preliminary clustering results as a way of illustrating the potential of this method to more definitively test the SINGULARITY assumption. In Figure 5, we show two dendrograms obtained by performing agglomerative clustering on each set of QUDs from one of our interviews, using Ward linkage and BERTScore as a Euclidean distance metric. In the first dendrogram, all subjects write very similar QUDs, similar enough that they are counted as a single cluster. In the second, there is more variability among QUDs: here the algorithm instead groups questions into several distinct, but internally similar clusters. For this latter set of QUDs, this analysis could be



(a) Single cluster

(b) Multiple clusters

Figure 5: (a) Highly similar QUDs arranged into a single cluster. (b) Dissimilar QUDs grouped into separate, internally similar clusters.

used to argue that, for that turn in discourse, there are potentially multiple contextually salient questions being addressed.

However, there are limitations to this method, too. While the ability to group questions into clusters is appealingly intuitive, clustering alone is not able to tell us whether the variability in inference arises because of inherent multiplicity or uncertainty — though further analysis of each cluster can give us better clues than depending on aggregate measures of variability. Also, as with many unsupervised methods, the number of clusters produced is highly dependent on hyperparameters like the linkage criterion and distance threshold, which prevents us from making any definitive claims about the exact number of QUDs being addressed. We leave overcoming these and other analytical challenges to future work.

6. Conclusion. Overall, we find that ordinary language users generally make similar inferences about the current Question Under Discussion, but there is considerable variability among subjects in the QUDs they produce. We interpret this data in light of three commonly held assumptions about QUDs in the literature: the SINGULARITY of the current QUD, Q-TO-QUD, and Q-A CONGRUENCE. Given the persistent variability, we believe that a realistic treatment of naturalistic discourse requires the relaxing our first assumption, SINGULARITY, though we stop short rejecting it outright. While some of this variability may be due to inherent multiplicity of contextually felicitous QUDs, some may be due to subject uncertainty or subject differences in how they integrate discourse context, and some may be due to limitations of our question similarity metrics. Our data suggest that our second assumption, Q-TO-QUD should also be relaxed, though more work on the nature and incidence of explicitly asked questions is needed to fully validate this suggestion. Lastly, we find that, despite a lack of explicit linguistic training, our subjects write QUDs which largely follow the assumption of Q-A CONGRUENCE. Through further collection and analysis of naturally-sourced QUDs, we hope to gain a better understanding of how QUD inference plays out in naturalistic discourse.

References

Cooper, Robin, Staffan Larsson, Elisabeth En-gdahl & Stina Ericsson. 2000. Accommodating questions and the nature of QUD .

De Kuthy, Kordula, Nils Reiter & 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).

Ginzburg, Jonathan. 1996. Dynamics and the semantics of dialogue. In Jerry Seligman, Dag Westerståhl & Lawrence Cavedon (eds.), *Logic, language, and computation*, vol. 1 CSLI Lecture Notes, 221–237. Stanford, California: CSLI Publications.

Halvani, Oren. 2024. Constituent Treelib - A Lightweight Python Library for Constructing, Processing, and Visualizing Constituent Trees. 10.5281/zenodo.10951644. `https://github.com/Halvani/constituent-treelib`.

Hamblin, C. L. 1973. Questions in Montague English. *Foundations of Language* 10(1). 41–53.

Majumder, Bodhisattwa Prasad, Shuyang Li, Jianmo Ni & Julian McAuley. 2020. Interview: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8129–8141. Online: Association for Computational Linguistics. 10.18653/v1/2020.emnlp-main.653.

Mulligan, Karl & Kyle Rawlins. 2024. Identifying Questions Under Discussion in Naturalistic Discourse. *Society for Computation in Linguistics* 7(1). 357–361. 10.7275/SCIL.2228.

Pyatkin, Valentina, Ayal Klein, Reut Tsarfaty & Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2804–2819. Online: Association for Computational Linguistics. 10.18653/v1/2020.emnlp-main.224.

Roberts, Craige. 1996/2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5. 10.3765/sp.5.6.

Rooth, Mats. 1992. A Theory of Focus Interpretation. *Natural Language Semantics* 1(1). 75–116.

Van Kuppevelt, Jan. 1995. Discourse Structure, Topicality and Questioning. *Journal of Linguistics*

31(1). 109–147.

Velleman, Leah & David Beaver. 2016. Question-based Models of Information Structure. In Caroline Féry & Shinichiro Ishihara (eds.), *The Oxford Handbook of Information Structure*, 0. Oxford University Press. 10.1093/oxfordhb/9780199642670.013.29.

Westera, Matthijs, Laia Mayol & Hannah Rohde. 2020. TED-Q: TED Talks and the Questions they Evoke. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1118–1127. Marseille, France: European Language Resources Association.

Wu, Yating, William Sheffield, Kyle Mahowald & Junyi Jessy Li. 2023. Elaborative Simplification as Implicit Questions Under Discussion. 10.48550/arXiv.2305.10387.

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

Appendix

Low question similarity example		
	QUD	Answer span
(L.1)	What can this hide?	[It can be used to deflect any missiles that might be based on radar return and it can hide aircraft.]
(L.2)	What are one of the uses of chaff in military operations?	[it can hide aircraft]
(L.3)	What is that in the sky?	[missiles]
(L.4)	Are there any other military applications?	[it can hide aircraft]
(L.5)	How can Chaff be used in other applications?	[It can be used to deflect any missiles that might be based on radar return and it can hide aircraft.]
(L.6)	How would a missile be affected by chaff?	[can be used to deflect]
Utterance: *It can be used to deflect any missiles that might be based on radar return and it can hide aircraft.*		

Table 2: A sample of low-similarity QUDs (mean BERTScore: 0.23).

High question similarity example		
	QUD	Answer span
(H.1)	Why were pregnant women scared about the Zika virus?	[their babies might be born with severe birth defects.]
(H.2)	Why are pregnant woman worried about the Zika virus?	[fear that their babies might be born with severe birth defects.]
(H.3)	What makes these women think their babies might be born with birth defects?	[women who've had the Zika virus]
(H.4)	What do the women from Brazil and Colombia fear?	[from Brazil and Colombia about pregnant women who've had the Zika virus and fear]
(H.5)	why are women in fear of the zika virus?	[their babies might be born with severe birth defects.]
(H.6)	What's concerning to pregnant women in Brazil and Columbia?	[pregnant women who've had the Zika virus and fear that their babies might be born with severe birth defects.]
Utterance: *This week, we've heard stories from Brazil and Colombia about pregnant women who've had the Zika virus and fear that their babies might be born with severe birth defects.*		

Table 3: A sample of high-similarity QUDs (mean BERTScore: 0.49).