Asking between the lines: Elicitation of evoked questions in text

Motivation. A concept that has become increasingly central to semantic/pragmatic accounts is Question Under Discussion (QUD; Carlson 1983; Kuppevelt 1995; Ginzburg & Sag 2000; Larsson 1998; Roberts 1996). In QUD accounts, discourse is structured via a sequence of discourse moves, comprising questions and their answers. Explicit QUDs influence both the surface form of the answer and the meaning derived from that answer. But not all QUDs are explicit, particularly in natural discourse, and implicit QUDs guide pragmatic interpretation for a range of phenomena ranging from implicature to presupposition to coreference (Cummins & Rohde 2016; Kehler & Rohde 2017).

Rather than modeling QUD recovery as a backward-looking derivation of discourse structure, or annotating QUDs with the entire discourse available (Riester 2019), here we target QUD *anticipation* (Onea 2016). We introduce a novel method for annotating potential and actual QUDs in transcribed speech, by assessing first what questions a discourse implicitly evokes and subsequently which of those are taken up as the discourse proceeds. We report metrics to validate the reliability of the data we elicited and outline ways in which the data can inform our understanding of QUD-driven phenomena and QUD models themselves. Our annotated texts, with color-coding of inter-annotator agreement measures, are available at https://evoque-data.github.io.

Method. Participants (N=87) read excerpts of transcribed speech, and were probed at multiple points in each text to type a question elicited by the utterances they had just read. The excerpts span three genres: two unscripted spoken dialogues (3807 words total, from Rehbein et al.'s (2016) DISCO-SPICE corpus), six scripted presentations (6975 words total, Zeyrek et al.'s (2018) TED-MDB) and one construct story (56 words) which we composed with certain obvious questions and answers in mind, as a sanity check for our method. At each probe, participants typed a question and highlighted a sequence of words in the text that primarily evoked that question (a highlighted 'topic'). Participants rated their question as 'answered' on a scale of 1 to 5 as they read the subsequent text, highlighting words that gave an answer. In total, 868 probe points elicited a mean of 5 questions each. The content of these questions and their highlighted topics and answers allow us to measure what words in a discourse evoke and address questions, and to compare such behavior across participants. The elicited questions which are rated high for 'answered' make for plausible QUDs, since those are anticipated questions which the speaker actually ends up addressing.

Data validation. Overall, participants clearly engaged with the task and posed questions which showed interest and anticipated the text's subsequent discourse moves. The resulting dataset contains 4765 questions with their highlighted topics and answers. Of these, less than half (around 2251) were rated to be 'not answered at all' (rating 1), with the rest quite evenly divided, around 650 each, between ratings 2, 3, 4 and 5 ('completely answered'); the mean 'answered' rating being 2.38. This shows that participants indeed ask questions that anticipate speakers' upcoming discourse moves, i.e., questions that are plausibly QUDs, although as expected there is also considerable indeterminacy.

As an example of a probe point with high 'answered' ratings (and considerable agreement between annotators, discussed shortly), the following example from a spoken dialogue elicited questions from 11 participants, all of whom highlighted the same 5-word topic and many of whom highlighted the same subsequent answer:

(1) `He was uh he was a bit upset on uh uhm first day the Friday` DISCO-SPICEp1a-094:line37

Elicited questions: *Why was he upset on his first day? Why was he upset? He was upset about what? Why was he upset? What happened to him? What happened to upset him? Is he better now? Why was he upset? Why was he upset? Why is he upset? Why was he upset?*

… `side effects of the medication` DISCO-SPICEp1a-094:line39

Genre	answered	SIF-sim	Same Q-type	# questions
DISCO-SPICE dialogue	2.11	.22	.19	2131
TED talk presentation	2.50	.27	.24	2412
Constructed story	2.89	.29	.27	222

Table 1: Inter-annotator agreement scores and 'answered' ratings for questions

We also looked at the distribution of elicited question *types*, defined essentially by the first word of the question barring some multi-word expressions (e.g., we analyze *how come* as the same type as *why*, not as *how). What*-questions were the most frequent, across genres, likely due to the flexibility of this wh-word. Auxiliary-initial polar questions were next, followed by *how*/*why*-questions. The DISCO-SPICE excerpts yielded more *who*/*where*-questions compared to TED talks and the constructed story, reflecting a higher proportion of clarifying and situating questions in these unscripted dialogues (e.g., *Who are they talking about? Where are they?).* Breakdown of 'answered' ratings by question type shows that the latter are also the least answered – our participants' meta/clarification questions were not as at-issue for the original interlocutors of the texts – while *why/how* questions were the most answered, suggesting more reliable anticipation of QUDs.

Breakdown of 'answered' ratings by genre is shown in Table 1, along with two scores for inter-annotator agreement. The 'answered' ratings show that, in line with expectation, anticipation of QUDs is easiest in our constructed story (as intended), hardest in unscripted dialogue, with TED talks in between. The scores for inter-annotator agreement are:

- **SIF-sim:** Cosine similarity of Smooth Inverse Frequency (SIF) sentence embeddings (Arora et al., 2017) of the elicited questions (a measure of semantic similarity). In a pilot study we annotated "question equivalence" by hand (e.g., all questions for (1) except "*Is he better now?"* would be assessed as equivalent), showing a Spearman correlation of 0.35 with the SIF-sim score. This, along with qualitative inspection of the data, suggests SIF-sim is reliable for the coarse-grained statistics reported here to validate our method.
- **Same Q-type:** Whether two questions are of the same question type (see above).

Table 1 reports these scores as an average over all probe points in the genre, where for each probe point we averaged the score over all pairs of questions elicited at that point. Compared to the unscripted DISCO-SPICE dialogues, we found that TED talks and the constructed story yielded questions with higher inter-annotator agreement, as expected.

Outlook. The dataset presented here represents the first batch of a larger scale collection, analysis, and ultimate release of a corpus annotated with potential and actual QUDs. Our hope is that this can serve as a benchmark dataset for evoked questions (akin to 'word association norms' and other existing human judgment benchmarks). In addition, already in its present form it offers a rich and multi-faceted window onto the way questions are anticipated and resolved in natural dialogue. We have shown only some coarse ways of validating the data we collected; mostly leaving out, for instance, which tokens our participants chose to highlight. As an example, we found that 'answered' scores are the highest by far when the highlighted words include a *wh*-pronoun, reflecting the fact that explicit questions in a discourse are more likely to end up as the next QUD (and then be answered). More analysis of our elicited data is ongoing, as well as alignment of these data with the discourse coherence relation annotations already available for DISCO-SPICE and TED-MDB. Through the latter we hope to be able to address interesting new questions about the relation between QUDs and coherence relations, two key notions in pragmatics, e.g.: Do coherence-signaling devices overlap with participants' highlighted topics and answers? Do speakers omit discourse connectives at a higher rate for utterances that answer more predictable QUDs? Manual inspection of the annotated texts remains valuable, too, and our browser-based visualization tool makes this easier: See https://evoque-data.github.io.

References.

Carlson, L. (1983). *Dialogue games: An approach to discourse analysis*. Dordrecht: Reidel.

Cummins, C. & Rohde, H. (2015). Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology*, 6(1779), 1-11.

Ginzburg, J., & Sag, I. (2000). *Interrogative investigations*. Stanford: CSLI Publications.

Kehler, A. & Rohde, H. (2017). Evaluating an expectation-driven QUD model of discourse interpretation. *Discourse Processes,* 54(3), 219-238.

Kuppevelt, J. van. (1995). Discourse structure, topicality, and questioning. *Journal of Linguistics*, 31, 109-147.

Larsson, S. (1998). Questions under discussion and dialogue moves. *TWENDIAL'98.*

Arora, S. and Y. Liang and T. Ma (2017). A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *International Conference on Learning Representations*.

Onea, E. (2016). *Potential questions at the semantics-pragmatics interface*. Brill.

Rehbein, I., Scholman, M. C. J., Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks*. LREC 16.*

Riester, A. (2019). Constructing QUD trees. In Questions in Discourse (pp. 164-193). Brill.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *OSU Working Papers in Linguistics*, 49: Papers in Semantics.

Zeyrek, D. Mendes, A. Kurfalı, M. (2018) Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank. *LREC2018.*