Asking between the lines: Elicitation of evoked questions in text

Matthijs Westera[1] and Hannah Rohde[2]

[1] Universitat Pompeu Fabra
matthijs.westera@gmail.com
[2] University of Edinburgh
hannah.rohde@ed.ac.uk

Abstract

We introduce a novel, scalable method aimed at annotating potential and actual Questions Under Discussion (QUDs) in naturalistic discourse. It consists of asking naive participants first what questions a certain portion of the discourse evokes for them and subsequently which of those end up being answered as the discourse proceeds. This paper outlines the method and design decisions that went into it and on characterizing high-level properties of the resulting data. We highlight ways in which the data gathered via our method could inform our understanding of QUD-driven phenomena and QUD models themselves. We also provide access to a visualization tool for viewing the evoked questions we gathered using this method (N=4765 from 111 crowdsourced annotators).

1 Introduction

A piece of discourse can evoke certain questions. For instance, the utterance "Latisha is worried." will quite reliably evoke, in an addressee's or other observer's mind, the question of why Latisha is worried – and an addressee is likely to respond accordingly: "Why?", "How come?", or "Worried about what?". The questions that a piece of discourse evokes provide a window into comprehenders' expectations about both the underlying situations being described by a text as well as the way in which the discourse itself is unfolding. They reflect what humans generally care about (e.g., causes of mental states) as well as what they understand or expect to be relevant in the particular discourse at hand, say, what they anticipate may be the next Question Under Discussion as intended by the original speaker (QUD [6, 21, 19, 13, 10]).

This paper summarizes our work-in-progress on annotating discourses with evoked questions using a crowdsourcing approach. More concretely, we present people snippets from a number of texts, and ask them to enter a question that the text evokes. Subsequently, we let them read how the text continues and ask them if their question has been answered. Moreover, we ask our participants to highlight the word or short phrase in the discourse which primarily evoked the question, and likewise for the words which primarily provided the answer. We repeat this process for multiple snippets per text to obtain evoked questions and, where available, their answers, after basically every sentence, in a number of texts. Altogether, this results in a rich dataset of evoked questions, their triggers in the text, and their associated answers.

This paper explains and motivates our data gathering method, presents some basic statistics on the resulting dataset as well as some coarse-grained analysis aimed merely at demonstrating the promise of our approach. Our main motivation for this undertaking is not the evoked questions as such – although we think they represent an important, hitherto unexploited cognitive/linguistic resource – but the window they provide on QUDs. We think of our method for eliciting evoked questions as the first step in a scaleable pipeline for annotating QUDs: After all, questions which were both evoked and subsequently answered are plausible candidates to be the original speaker's intended QUD. Let us briefly clarify the need for a scalable QUD-annotation pipeline: QUDs form the backbone of many current semantic/pragmatic theories. Such QUD-based theories often make falsifiable predictions only once the assumption of a certain type of QUD (or a certain set of 'alternatives') is granted (just as the Gricean maxims [11] presupposed a notion of 'relevance'). Now, this isn't specifically problematic when merely assessing such theories intuitively using constructed linguistic examples, or when testing them in a controlled empirical experiment with constructed stimuli, since in both cases the relevant linguistic examples can be presented against the background of an explicit interrogative sentence, thereby fixing the intended QUD. But in naturally occurring discourse explicit interrogatives are rare, rendering this domain primarily off-limits for large-scale hypothesis testing of QUD-based theories. Clearly this situation is unsatisfactory. The remedy would be a powerful, quantitative model of QUDs that can tell us, for any stretch of natural discourse, which QUDs are plausible there and which ones aren't. In order to develop (train, test) such a model, we will need a large amount of naturalistic data annotated with QUDs. For this reason it is important that we come up with a scalable QUD annotation pipeline, one which can be crowdsourced rather than relying on expert annotators (who are expensive, scarce, and potentially theoretically biased).

Although our main aim is to work towards a scalable QUD-annotation pipeline, we also think, as we mentioned, that the notion of an evoked question is interesting in its own right. Just as benchmark datasets for something like word similarity or word association are frequently relied upon (e.g., given the word "snow", write the first alternative words that come to mind), we think that a benchmark of evoked questions (given a piece of discourse, write the first question that comes to mind) will provide an important window on linguistic competence and on the human mind more generally. Where word similarity benchmarks have become indispensable in lexical semantics (especially computational approaches), we think an evoked questions benchmark could become indispensable for pragmatics.

2 Related work

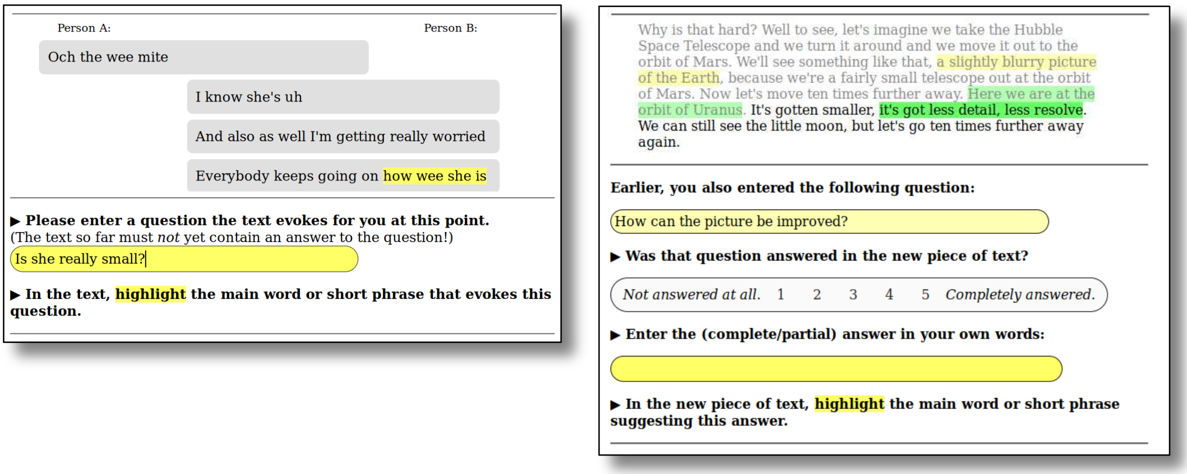
Questions Under Discussion (QUDs) are frequently relied upon in semantics and pragmatics in theories of discourse structure and the wide range of phenomena that depend on it, such as implicature (e.g., [9, 14, 2, 23]), discourse particles (e.g., [20]), prosody (e.g., [19, 5, 3, 24]) and pronoun interpretation (e.g., [7, 12]). However work on annotating QUDs is limited: to our awareness the only attempt is [8, 18]. Compared to that approach, our annotation task is much more free: our participants do not receive detailed instructions and can basically enter any question they like, whereas [8, 18] carefully formulated a number of (theory-derived) constraints on QUDs, such as the constraint that the QUDs of a discourse should form a tree. As a consequence, whereas they relied on instructed expert annotators, our procedure is sufficiently simple and natural for naive annotators, hence for a scalable crowdsourcing approach – though of course whether annotations obtained from the two different procedures share some features of interest remains to be seen. Another major difference is that their annotators were given the full discourse, whereas for our participants the texts are revealed one snippet at a time, at each point elicitating an evoked question (and asking if previous questions have been answered yet), prior to revealing how the text continues. One reason for this is that considering which questions (potential QUDs) a discourse *evokes* is a simpler, far more natural task than deciding which question an utterance *addresses* (actual QUD), and therefore that it lends itself better to crowdsourcing. Another reason is our interest in the forward-looking process itself, i.e., the anticipation/predictability of QUDs, drawing inspiration from [15].

While QUD annotation has only recently been undertaken, and on limited data, the annotation of discourse relations has a long history: e.g., the Penn Discourse TreeBank (PDTB; [16]). Contrary to QUD-theory – where the range of possible QUDs is in principle open-ended – the PDTB and other coherence annotation schemes recognize only a limited number of types of discourse relations. Moreover, for the PDTB the different types of relations are tied to particular natural language connectives such as “and then” and “because”, such that annotating discourse relations is a matter of (i) recognizing and categorizing explicit connectives and (ii) recognizing places where such connectives are arguably implicit, say, where they could naturally be inserted. This setup greatly facilitates annotation, though at the cost of relying on an a priori decided, fixed repertoire of relation types. It will be interesting to compare existing discourse relation annotations to the evoked questions we elicit – one would of course expect some kind of alignment, e.g., a ‘causal’ discourse relation may tend to occur after snippets that evoke questions such as “Why?” and “How come?”. To enable this kind of research we decided to elicit evoked questions for two small datasets that have already been annotated in PDTB style: TED-MDB (Multilingual Discourse Bank, six TED talk transcripts in different languages, though we’ll be using only the English portion; [26]) and DISCO-SPICE (dialogue and interview transcripts in Irish English; [17]). Although in the present paper we will not yet look at the alignment of discourse relations and evoked questions, we have recently conducted this research and report on it in a manuscript under review [25].

QUDs are variably regarded as a type of discourse move or as a representation of the discourse goal, though often a mix of both. The first perspective regards QUDs as a questioning speech act which can be either explicit (i.e., realized by an interrogative sentence in the discourse) or implicit. The second perspective regards them as a model of discourse goals, and goals are strictly speaking different creatures from the speech acts intended to serve them; both questioning and asserting speech acts must relate to the discourse goal, but are conceptually distinct from this goal, and also formally not necessarily equivalent to it. We think the second perspective is ultimately the more adequate/fruitful one: we need a notion of discourse goals in our theories, and they must be distinguished from the acts by means of which speakers try to achieve them. Nevertheless, the two perspectives are often quite harmlessly conflated: the main purpose of questioning speech acts, and of the interrogative sentences that express them, is to set the discourse goal for the next speaker; and any discourse goal can be fairly naturally paraphrased by means of an interrogative. We too rely on interrogative sentences – namely, the evoked questions we elicit from our participants – as a natural way of annotating the potential and actual QUDs of the discourse. But we must remain aware that this is a simplification, as interrogative sentences often explicate only a part of their QUD: just as a declarative utterance typically asserts only some of the propositions in the QUD, so too may an interrogative utterance highlight some of them while backgrounding others (e.g., [4, 22]).

3 Method

For concreteness we will first show what our elicitation task looked like, before going into details about how we cut up the texts to serve them in portions to our participants. The core of our method is to present participants a piece of text, ask them which question it evokes, then reveal how the text continues and ask whether their question has been answered. We set up our task in Ibex (Internet-based experiments, https://github.com/addrummond/ibex), although our incremental (chunk-wise) display of text and giving participants the ability to highlight text required a lot of customization. The two main screens of the task are shown in figure 1. As can be seen in the left figure, we simply asked participants to “enter a question the text evokes at this point”, and (motivated by pilot results) we included a clarification that it was to be



Person A:

Person B:

Och the wee mite

I know she's uh

And also as well I'm getting really worried

Everybody keeps going on how wee she is

▶ **Please enter a question the text evokes for you at this point.**
(The text so far must *not* yet contain an answer to the question!)

Is she really small?

▶ **In the text, highlight the main word or short phrase that evokes this question.**

Why is that hard? Well to see, let's imagine we take the Hubble Space Telescope and we turn it around and we move it out to the orbit of Mars. We'll see something like that, a slightly blurry picture of the Earth, because we're a fairly small telescope out at the orbit of Mars. Now let's move ten times further away. Here we are at the orbit of Uranus. It's gotten smaller, it's got less detail, less resolve. We can still see the little moon, but let's go ten times further away again.

Earlier, you also entered the following question:

How can the picture be improved?

▶ **Was that question answered in the new piece of text?**

Not answered at all. 1 2 3 4 5 *Completely answered.*

▶ **Enter the (complete/partial) answer in your own words:**

▶ **In the new piece of text, highlight the main word or short phrase suggesting this answer.**

Figure 1: Our elicitation tool, asking for an evoked question (left) and whether a prior question has been answered (right); dialogue (left) and monologue (right) are rendered differently.

a question that hadn't yet been answered at that point. Other than this we did not provide instructions about what counts as an 'evoked question', trusting that it would be a sufficiently natural task. In particular, we did not instruct them to try to guess where the discourse was heading, although we do believe that our task biases them towards this, by repeatedly asking whether a question has been answered yet.

After a small pilot study which resulted in very few elicited questions, we decided to make entering a question mandatory, because we wanted our workers to make an effort, and there's always *something* that can be asked. (Although we may be 'overeliciting' questions in this way, questions that are too random/too desperate could in principle be filtered out by looking at inter-annotator agreement; cf. section 4.) After eliciting an evoked question, we would reveal the next chunk of text and ask them whether their question was answered at that point. This is shown in figure 1 on the right (though for a different text than the one on the left): we repeated the question they entered and asked for a rating on a 5-point Scale. For both questions and answers, participants were asked to highlight the main word or short phase in the text that primarily evoked the question, or provided the answer, respectively. They could do this by dragging a selection in the newest chunk of text, where they could highlight a single span of up to 10 words. Highlights for different questions were given different colors, and highlights for answers were given the same color as the question they were answers too, which was intended as an additional visual guide to decrease cognitive load. We would repeat this procedure, incrementally taking each participant through a text, at each point asking for evoked questions as well as possible answers to questions evoked earlier.

We hosted our Ibex experiment on IbexFarm (`http://spellout.net/ibexfarm/`), and recruited participants from Amazon Mechanical Turk (`http://mturk.com`). Each participant could do the task once, where a task would consist of 6 excerpts, where each excerpt is comprised of 8 chunks of text (i.e., 8 probe points asking them for questions/answers). We estimated that the task would take about 50 minutes, and offered a compensation of $8.50. Comments from our workers suggest that the task was a bit long and became boring after a while, suggesting that we should have offered perhaps only one excerpt per task – this is something we'd do differently in the future when getting additional annotations for the same texts and/or annotating different texts. We aimed to have at least 5 participants for every probe point, but because we let the excerpts overlap many probe points have more than that.

As for the texts we used, as mentioned earlier we rely on two existing datasets, to which we contribute the elicited evoked questions as an additional layer of annotations. These datasets are TED-MDB [26] and DISCO-SPICE [17]. TED-MDB consists of transcripts of six scripted presentations from the TED Talks franchise, in multiple languages, from which we use the English portion (6975 words total). DISCO-SPICE consists of transcripts of Irish conversational English, of which at present we used only two transcripts (3807 words total; but comparable to TED-MDB in number of sentences). In addition we included a simple short story (56 words), which we ourselves composed with certain obvious evoked questions and subsequent answers in mind, as an introductory item for our participants and as a sanity check for our method.

Our aim was get full coverage of each text, namely, to elicit evoked questions and check for answers either after every sentence or at certain points within long sentences. To that end we first cut each text into sentences (or sentence-like fragments). For DISCO-SPICE we relied on the existing division into utterances; for TED-MDB we ran NLTK's sentence tokenizer. Any remaining long sentences ($>$ 150 words) were further automatically cut up (as necessary) at commas, colons or semicolons, both in order to elicit questions at those points, and to make all chunks more or less the same length. We also manually inserted cuts just prior to occurrences of "because", anticipating that it could be interesting to investigate which questions are evoked there. For convenience we will refer to the resulting sentence-like pieces of text as sentences.

Although we wanted to elicit evoked questions at every sentence, unfolding the discourse one sentence at a time and asking participants to enter a question at each point will get tedious quickly and break the flow of the discourse – and often a single sentence will not contain much new information anyway to evoke a meaningful question or provide an answer. Therefore, we decided to unfold each excerpt in chunks of 2 sentences each, and probe each participant only after every such chunk. We rotated participants through the different ways of dividing a text into chunks to get full coverage this way. Altogether, this resulted in 460 probe points covering the six texts of TED-MDB and 397 probe points covering the two texts of DISCO-SPICE.

We decided not to present full texts to our participants, but only excerpts of around 10 of the aforementioned chunks, before letting them move on to an excerpt from a different text. This decision aligned with our aim to make our elicitation process part of a scalable pipeline for QUD-annotation, a pipeline which should be able in principle to target very long texts – texts too long for a single participant to cover in a timespan reasonable for crowdsourcing. It does come with a downside, namely, that we will be able to capture only relatively 'local' discourse structure. But we think that eliciting evoked questions through crowdsourcing is already biased towards local discourse structure to begin with, by virtue of the fact that coming up with local questions and finding their answers is an easier way of completing the task than asking high-level questions, which would require keeping the whole text in mind. We think that capturing more 'global' discourse structure would require a different type of task.

With the choice to give participants only excerpts instead of full texts, there is a risk that at the start of an excerpt they will lack the context necessary to understand what's going on, such that the evoked questions will mostly be clarification questions ("who are the speakers?", "what are they talking about?") that do not reflect the actual discourse structure for those who were present. For this reason we let the excerpts of a given text overlap, such that what are the first chunks for some participants (who may lack context) will be chunks at the end for other participants. We can then in principle check whether there are systematic systematic between the questions evoked after excerpt-initial and excerpt-final chunks and, if necessary, discard excerpt-initial chunks without loosing coverage of those pieces of discourse altogether.

Another decision that bears on whether we get to capture local or global structure is the following. Recall that, in addition to asking for a question evoked by the text, we ask our participants whether their previous questions have been answered yet. This raises the question of how long we should keep pestering workers with questions they entered that remain unanswered: if we keep at this for too long then the task will get tedious very quickly – "is it answered yet?", "and now?", "what about now?" – especially given the fact that most evoked questions will never be answered. We conducted a pilot where we prompted participants with the questions they entered for up to three chunks after the chunk that evoked them. We found that questions that didn't get answered after one chunk were practically never answered within three chunks, either. Therefore, to cut task duration (hence cost) without risking substantial data loss we decided to decrease this limit to two: participants are asked at most twice (i.e., after two subsequent chunks of text) whether any given question they entered had been answered. This also meant that the highest number of unanswered questions a participant ever had to consider at any particular chunk was two. In any case, this is another respect in which our approach is biased towards local discourse structure; it is not suitable for getting at questions that are answered only after a longer stretch of discourse unless those questions are re-evoked at a subsequent probe point.

To summarize: we cut the source texts into overlapping excerpts of up to 8 chunks each, where each chunk is made up of two sentence-like parts (getting full coverage by rotating our participants through the different ways of cutting each text into chunks). During the task, unanswered evoked questions were kept around for up to two chunks. Our procedure is biased towards eliciting questions that reflect 'local' QUD structure. For the intended comparison to the discourse relation annotations, which we leave to [25], this restriction to local structure is quite alright: discourse relations are likewise predominantly local, since they are based around explicit or implicit connectives between clauses. But ultimately a complementary task would be required to get at the more global structure.

4 Bird's-eye overview of the data

Although our focus in this paper is on the method and motivations for our design decisions, we give a very general overview of the data collected. For qualitative, manual inspection of the resulting data we have published the texts annotated with the evoked questions we elicited, with color-coding of inter-annotator agreement as discussed below, at https://evoque-data.github.io. The full dataset, including answers and highlights, and properly (re-)aligned with the existing discourse relation annotations, will be made available at a later date.

We recruited 111 participants from Amazon Mechanical Turk, who gave us a total of 4765 evoked questions for our 863 probe points. These and other numbers are summarized in Table 1. Overall, participants seemed to engage with the task and posed questions which showed interest and anticipated the text's subsequent discourse moves. We compute an Answered rating for evoked questions as the highest Answered rating (on a scale from 1 to 5) assigned to that question in the two chunks following it, i.e., the quality of its best answer. Computed thus, the mean Answered rating over all questions is 2.38. Less than half of the evoked questions (2251) were rated to be 'not answered at all' (rating 1), with the rest quite evenly divided between ratings 2, 3, 4 and 5 ('completely answered'). So, participants do ask questions that anticipate speakers' upcoming discourse moves, i.e., questions that end up being at least partially answered hence are plausibly QUDs, but also many questions that don't get resolved at all.

Source texts:	8	Participants:	111	Probe points:	863
Word count:	10782	Excerpts/participant:	6	Questions:	4765
Excerpts:	142	Probe points/excerpt:	8	Answers:	1965
		Participants/probe point:	≥ 5		

Table 1: Summary statistics (excluding the constructed story all participants saw as practice).

Genre	ANSWERED	SIF-SIMILARITY	Same question 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 2: Inter-annotator agreement scores and 'answered' ratings for questions

As an example of a probe point with high 'answered' ratings, the following example from a spoken dialogue elicited questions from 11 participants:

(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?
Highlighted answer:
[...] The oul side-effects of the medication. DISCO-SPICEp1a-094:line39

Besides most questions elicited for this chunk indeed being answered, there is also high agreement among our participants about what question was evoked. Moreover, all of them highlighted the same five-word topic ("he was a bit upset") as evoking the question, and most of those whose questions were answered highlighted the same subsequent answer. We will attempt to quantify inter-annotator agreement shortly below.

First let us consider what types of questions were evoked. Here we define the 'question type' essentially by the first word of the question (wh-word or auxiliary verb), although we also took some multi-word expressions into account; for instance, we treat "how come" as the same type as "why", not as "how". Based on our coarse-grained classification of question types, *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 (along with a measure of inter-annotator agreement, see below) by question type is shown in figure 2. It shows that the *who*/*where*-questions – many of which are clarifying/situating questions – are also the least answered, as expeceted: these meta questions were not 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 the first column of Table 2. 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 other columns in Table 2 represent two coarse-grained measures of inter-annotator agreement,

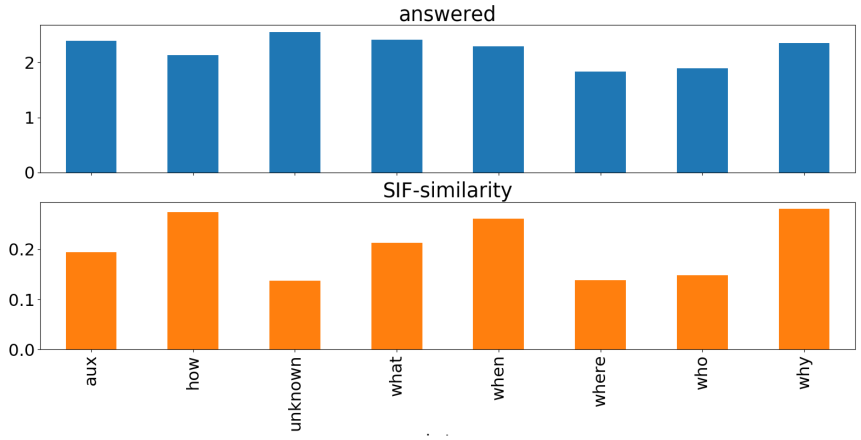


Figure 2: Answered rating and SIF-similarity (inter-annotator agreement) by question type.

and '# questions' the total number of questions in each subset of the data. The 'same question type' column represents simply the average proportion of pairs of questions (evoked by a given chunk) that are of the same question type as defined above, i.e., based simply on the first token(s) of the questions. The SIF-similarity column represents the average of a notion of semantic similarity computed between all pairs of questions evoked by a given chunk. It is computed as the cosine similarity of Smooth Inverse Frequency (SIF) sentence embeddings ([1]) of the elicited questions, which is a measure of (composed) distributional semantic similarity. Both measures of inter-annotator agreement show that TED talks and the constructed story yielded questions with higher inter-annotator agreement compared to the unscripted DISCO-SPICE dialogues, as expected, and in line with the pattern exhibited by the Answered scores.

A quick note about the reliability of SIF-similarity is in order. In a smaller pilot study we ourselves annotated 'question equivalence' by hand (e.g., all questions for (1) except "Is he better now?" would be assessed as equivalent). This resulted in a Spearman correlation of 0.35 between our annotations and the SIF-similarity score, suggesting that the latter may indeed be sufficiently reliable for the coarse-grained statistics reported here, but also that a more accurate (more human) assessment of question equivalence, or inter-annotator agreement more generally, is clearly called for. In particular, the kinds of question pairs that SIF-similarity misclassifies as unrelated are ones that happen to use very different words, or cases where one question is very explicit and the other is relying more on context (pronouns, ellipsis). To remedy this, we have recently fed our evoked questions through another round of crowdsourcing, where we give workers a snippet with all the questions it evoked and ask them to indicate how related the different pairs of questions are ('equivalence' being one extreme). We describe this in a manuscript currently under review [25].

We have shown only some very coarse ways of looking at the data we collected. We have hardly looked, for instance, at which tokens our participants chose to highlight. Just to give an example of the detailed findings this rich data supports: 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 very likely to end up as the next QUD (and then be answered). Lastly, manual inspection of the annotated texts remains valuable, too, and our simple online visualization makes this easier: See https://evoque-data.github.io.

5 Outlook

We have introduced a novel, scalable method aimed at annotating potential and actual Questions Under Discussion (QUDs) in naturalistic discourse. It consists in asking naive participants first what questions a certain portion of the discourse evokes for them and subsequently which of those end up being answered as the discourse proceeds. Our focus in this paper has been on outlining the method and design decisions that went into it, and on exploring some high-level properties of the resulting data with the modest hope of showing the promise of this approach.

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; analysis in this respect of the latter is reported in a paper under review [25]. This was one of our main reasons for choosing these two datasets: adding an evoked questions tier to the existing discourse relation annotations allows us 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?

The data we elicited for the present paper 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 will be a valuable tool for developing and testing both semantic/pragmatic theories and computational models of QUDs. In addition, we think that the notion of an evoked question is cognitively/linguistically interesting in its own right, independently of their relation to QUDs, and we hope that our data can serve as an 'evoked questions benchmark', perhaps akin to 'word association norms' and other existing human judgment benchmarks.

6 Acknowledgments

This work was supported in part by a Leverhulme Trust Prize in Languages and Literatures to H. Rohde. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

References

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*. 2017.

[2] David Beaver, Mandy Simons, Craige Roberts, and Judith Tonhauser. Questions under discussion: Where information structure meets projective content. *Annual Review of Linguistics*, 3:265–284, 2017.

[3] David I. Beaver and Brady Z. Clark. *Sense and Sensitivity: How Focus Determines Meaning*. Wiley-Blackwell, West Sussex, UK, 2008.

[4] María Biezma and Kyle Rawlins. Responding to alternative and polar questions. *Linguistics and Philosophy*, 35(5):361–406, 2012.

[5] Daniel Buring. On d-trees, beans, and b-accents. *Linguistics and Philosophy*, 26:511–545, 2003.

[6] Lauri Carlson. *Dialogue Games: An Approach to Discourse Analysis*. Reidel, Dordrecht, 1983.

[7] Chris Cummins and Hannah Rohde. Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology, Special issue on Context in communication: A cognitive view*, 6:1–11, 2015.

[8] Kordula De Kuthy, Nils Reiter, and Arndt Riester. QUD-based annotation of discourse structure and information structure: Tool and evaluation. In Nicoletta Calzolari et al., editor, *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, pages 1932–1938, 2018.

[9] Judith Degen. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics & Pragmatics*, 8,article 11:1–55, 2015.

[10] Jonathan Ginzburg and Ivan Sag. *Interrogative Investigations.* CSLI Publications, Stanford, 2000.

[11] H. P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics, vol. 3: Speech Acts*, pages 41–58. Academic Press, 1975.

[12] Andrew Kehler and Hannah Rohde. Evaluating an expectation-driven QUD model of discourse interpretation. *Discourse Processes*, 54:219–238, 2017.

[13] Staffan Larsson. Questions under discussion and dialogue moves. In *Proceedings of TWLT 13/Twendial '98: Formal Semantics and Pragmatics of Dialogue*, 1998.

[14] Laia Mayol and Elena Castroviejo. Projective meaning and implicature cancellation. In G. Kierstead, editor, *Proceedings of the ESSLLI 2011 Workshop on Projective Meaning*, 2011.

[15] Edgar Onea. *Potential questions at the semantics-pragmatics interface.* Brill, 2016.

[16] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2961–2968, 2008.

[17] Ines Rehbein, Merel Scholman, and Vera Demberg. Disco-spice (spoken conversations from the spice-ireland corpus annotated with discourse relations). In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 16)*, 2016.

[18] Arndt Riester. Constructing QUD trees. In Klaus v. Heusinger, Edgar Onea, and Malte Zimmermann, editors, *Questions in Discourse*, volume 2. Brill, Leiden, 2019.

[19] Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. *OSU Working Papers in Linguistics*, 49: Papers in Semantics, 1996.

[20] Tania Rojas-Esponda. A QUD account of german doch. In *Proceedings of the 18th Sinn und Bedeutung*, page 35976, 2014.

[21] Jan van Kuppevelt. Discourse structure, topicality, and questioning. *Journal of Linguistics*, 31:109–147, 1995.

[22] Matthijs Westera. *Exhaustivity and intonation: a unified theory.* PhD thesis, University of Amsterdam, 2017.

[23] Matthijs Westera. An attention-based explanation for some exhaustivity operators. In *Proceedings of Sinn und Bedeutung*, volume 21, pages 1307–1324, 2018.

[24] Matthijs Westera. Rise-fall-rise as a marker of secondary QUDs. In Daniel Gutzmann and Katharina Turgay, editors, *Secondary content: The semantics and pragmatics of side issues.* Brill, Leiden, 2019.

[25] Matthijs Westera, Laia Mayol, and Hannah Rohde. TED-Q: a dataset of TED talks with the Questions they evoke. In , under review.

[26] Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. TED multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style. In *Proceedings of the 11th Language Resources and Evaluation (LREC)*, 2018.