What are you talking about? Estimating the probability of Questions Under Discussion based on crowdsourced non-expert annotations

1 Goal and motivation

In discourse models based on Questions Under Discussion (QUDs) (von Stutterheim & Klein, 1989; van Kuppevelt, 1995; Roberts, 2012), it is assumed that each assertion in a discourse answers one such QUD. These QUDs can be annotated in natural texts, which leads to potentially hierarchical models of discourse structure on different levels of granularity. In the existing qualitative approaches, trained annotators assess what QUDs an assertion answers based on elaborate annotation guidelines (De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019). However, an assertion like (1a) can potentially be the response to a set of different QUDs, like (1b), and comprehenders are not necessarily certain about which of these questions the assertion answers. This uncertainty can be represented by a probability distribution over potential QUDs for each assertion, and the degree of uncertainty in this distribution might vary between assertions. In our contribution to the QUD-Anno Challenge, we present a data set collected in order to investigate (i) which QUDs comprehenders actually assume when processing the annotation materials of the challenge and (ii) to what extent the distribution over possible QUDs varies across the course of the text. Additionally, (iii) the data set might also complement the more fine-grained expert annotations provided by other contributors to the challenge.

(1) a. Susie lifts the lid of the abandoned teapot […].
b. {What is happening?, What does Susie do?, What does Susie lift the lid of?, …}

2 Relationship to previous approaches

While most previous empirical approaches to QUD-based discourse structure are qualitative (De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019), we take a quantitative approach and ask participants in a production task which QUD they think that an assertion like (1a) in the given context actually answers. After having preprocessed their responses, we obtain an approximation toward the set of possible QUDs for each assertion in a text and a probability distribution over these QUDs for each assertion. This approach is similar to the QUD production tasks by Westera et al. (2020), Poppels and Kehler (to appear) and Reich et al. (to appear). However, Westera et al. (2020) asked subjects to produce QUDs evoked by the text *up to a particular point* rather than asking which question a particular utterance probably answers. The other two studies collected QUDs only for isolated utterances, but not for a series of related assertions within a coherent text. To our knowledge, there is no quantitative research on the distribution of QUDs that an assertion answers in the course of larger texts.

3 Data collection

3.1 Method and participants

We conducted an online production experiment with 61 naive English speaking subjects recruited on the crowdsourcing platform *Prolific*, who are very likely not to have experience in the QUD-based annotation of discourse structure. Since the primary goal of our data collection was to explore the viability and outcome of our quantitative approach to collecting QUDs for a larger text, we collected QUDs for only the beginnings of two of the annotation material texts -– the narrative text and the car review. The narrative text was worked on by 30, the car review

by 31 self-reported native speakers of British English between the ages of 18 and 40.[1] They received a compensation of 2.60 £ for the shorter narrative text and of 3.50 £ for the longer car review.

3.2 Materials

The two texts were segmented into atomic assertions similarly to the strategy proposed by De Kuthy et al. (2018) according to the following criteria: Each declarative utterance, as delimited by periods, colons and semicolons was considered an assertion (including those being fragments, Morgan, 1973). Clausal and verbal coordinations were split into conjuncts, with each conjunct being considered an independent assertion (see the segmentation example in (2)), where assertions are labeled "A"), because these conjuncts (can) answer independent QUDs.

(2) [Susie lifts the lid of the abandoned teapot]$_{A}$ and [swirls the water]$_{A}$

We tested the initial 20 assertions from the narrative text and the initial 19 assertions from the car review.[2] For non-declarative utterances (e.g. interrogatives and imperatives occurring in the car review text), we did not collect QUDs, but since they also determine the text's discourse structure, they were shown to participants where they appeared in the texts.

3.3 Procedure

We implemented one survey for each text on the web-based experimentation platform *PCIbex* (Zehr & Schwarz, 2018). In the instructions, the participants were introduced to the idea that a declarative utterance can be understood as answering a potentially implicit question. They were also informed that a single utterance can answer different questions, and that not only complete sentences but also parts of sentences (like the above-mentioned conjuncts of clausal and verbal coordinations) can answer individual questions. We provided examples for this which were not related to the tested data (e.g. *Sue and Ann went to an Italian restaurant* as an utterance that might answer questions like *What happened?* or *Where did Mary and Ann go?*).

The participants' task consisted in entering the question that was answered by each of the assertions into a text field. They were instructed to enter only one question per assertion, to provide the one they considered most likely, and to restrain themselves from being funny or creative. Each participant read one of the two texts. Before reading the first assertion, participants saw the title and author of the text, because "normal" readers of these texts also have this information and can adapt their expectations about discourse structure accordingly. The text itself was then incrementally revealed assertion by assertion, with the new assertion marked in blue font and appearing below the rest of the text processed so far, which was displayed in black font. A demo version of the experiment can be found here: https://farm.pcibex.net/r/lBOAcB/

4 Data set creation

The goal of the annotation procedure was to pool, for each assertion, the different, but semantically identical lexicalizations of each QUD. Otherwise, the probability mass of a single QUD would be split between synonymous expressions in the probability distribution over QUDs.

[1] We had data from 31 instead of the anticipated 30 participants for the car review because one participant apparently did not enter their completion code on the *Prolific* website. For the narrative text, we excluded two participants who had not produced meaningful responses and replaced them by two new participants to end up with 30 participants.

[2] We stopped at 19 assertions for the car review because the next assertion was part of a complex sentence and it would have been odd to present only a part of it.

QUD	n	rank	probability
What did Susie do?	11	1	0.37
What did Susie do to the teapot?	6	2	0.2
What did Susie lift?	4	3	0.13
What did Susie lift the lid of?	2	4	0.07
Is there anything inside this teapot?	1	5	0.03
What did Susie do next?	1	5	0.03
…	…	…	…

Table 1: Section of the probability distribution for the first assertion of the narrative text.

4.1 Preprocessing and exclusions

Before the annotation, we excluded non-meaningful responses which resulted in a loss of 6.02% of the data (5.8% in the narrative, 6.23% in the car review). This concerned the copying of (parts of) the utterance into the text field, questions related to the task rather than the text (e.g. *What have I missed?*) and responses that were declarative statements or bare DPs instead of questions (e.g. *Tell me some details about Susie's life* or *waiting time*). Subordinate questions lacking a matrix clause (e.g., *How the tea was*) remained in the data set since such questions are essentially semantically equivalent to independent questions. If participants produced more than one question, we entered each of the questions separately into the data set. The final data set consists of 568 questions for the 20 assertions from the narrative text and of 557 questions for the 19 assertions from the car review.

4.2 Annotation

In the annotation process, all semantically identical QUDs produced for a single assertion were assigned a single unique label. Semantic identity was operationalized as having the same set of answer propositions (Hamblin, 1973; Karttunen, 1977). For example, questions like *What does Susie do?* and *What is Susie doing?* were grouped together. On the other hand, questions like *What did Susie do?* and *What did Susie do to the teapot?* received different labels, because the PP argument in the latter question (i.e. *to the teapot*) further restricts the set of possible answers. The most frequent lexical realization of a QUD, corrected for spelling, was used as a label for this QUD. The annotation was performed by two expert annotators (Schäfer and Hristova), who agreed on a single gold standard. The participants' individual responses and the assigned labels can be found in the attached csv-files.

5 Structure of the data set

The annotated data set consists in a set of QUDs for each of the assertions in the texts. For each QUD we provide its probability

$$p(QUD_i) = \frac{n(\text{QUD}_i)}{\sum_{i' \in \text{ QUD-SET}} n(\text{QUD}_{i'})} \tag{1}$$

and its frequency rank with respect to the other QUDs within its set. For example, the QUD set for the assertion *Susie lifts the lid of the abandoned teapot* has the probability distribution shown in Table 1.

For each assertion, we calculated the entropy H in the probability distribution over QUDs in its QUD set, which is a measure of the degree of uncertainty about the outcome of a random variable (Shannon, 1948, p. 393) as shown in equation 2. Entropy is maximal if all of the QUDs in a set are equally likely, and it equals 0 if there is only one QUD.

$$H = -\sum_{i=1}^{n} p_i log_2 p_i \quad (2)$$

The data are provided in an XML schema. Each QUD is represented through a QUD element, which contains the assertion it answers within a UNIT element.[3] Each QUD element has several attributes: The QUD label as "string", the classification as QUD, the identifier of the utterance which answers it as "utterance-id", a unique identifier for each QUD as "qud-sub-id", the QUD's frequency rank as "rank" and its probability as "probability". For each UNIT element, we list the numeric ID of the assertion as "utterance-id" and the entropy in this assertion's QUD as "entropy". The structure is illustrated in Figure 1.

```
<QUD string="What did Susie do?" classification="QUD"
utterance-id="1" qud-sub-id="1" rank="1" probability=
"0.37" >
    <UNIT utterance-id="1" entropy="2.79">Susie lifts
    the lid of the abandoned teapot</UNIT>
</QUD>
<QUD string="What did Susie do to the teapot?"
classification="QUD" utterance-id="1" qud-sub-id="2"
rank="2" probability="0.2" >
    <UNIT utterance-id="1" entropy="2.79">Susie lifts
    the lid of the abandoned teapot</UNIT>
</QUD>
```

Figure 1: The XML structure illustrated at the beginning of the narrative text.

6 Data set statistics

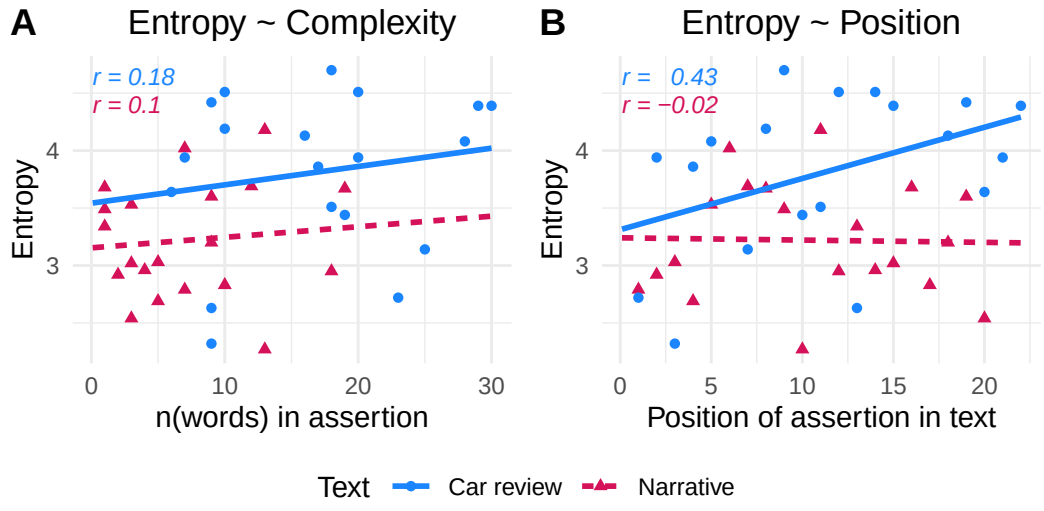

Figure 2: Correlation of the entropy with (**A**) the complexity of an assertion (measured in number of words) and (**B**) the position of the assertion in a text as a function of the text. Points correspond to individual assertions.

[3]This results in the redundancy that each assertion occurs several times in the data set, but it ensures that the XML structure is comparable to the qualitative expert annotations in that each QUD dominates the assertion it is answered by.

6.1 Variation between QUDs

Overall, for most assertions there is clearly one QUD which is more often produced than the other ones: The QUDs on rank 1 for each assertion have a mean probability of 0.28 (sd = 0.12), whereas the QUDs with rank 2 only have a mean probability of 0.13 (sd = 0.06). This however also indicates that there is considerable variation between QUDs: On average, 72% of the total probability mass is split among the less likely QUDs and in both texts, at least 8 different QUDs were produced for each assertion. For example, *What did Susie do?* is clearly the most probable QUD for the first assertion of the narrative. But since it has a probability of 0.37, this means that still for more than 60% of the cases, participants produced an different QUD than the most probable one. This is particularly noteworthy since we asked subjects to produce only one most likely QUD, which might skew the participants' underlying probability distribution toward the most likely one: If a QUD is actually on rank 2 for each participant, it will never be produced in spite of its relatively high probability. Taken together, the variation between the QUDs produced suggests that a quantitative approach to the investigation of QUD-based discourse structure is necessary to account for the complete empirical picture of the distribution of QUDs across a text.

6.2 Differences between text types

We also observed some tentative differences between the two text types for which we collected data (narrative and car review) with respect to the distribution of QUDs. First, the difference between the first and second ranked QUD differs between both texts: It is larger for the narrative text (rank 1 = 0.34 (0.13) vs. rank 2 = 0.13 (0.06)) than for the car review (rank 1 = 0.21 (0.12) vs. rank 2 = 0.12 (0.06)), which indicates that participants agreed overall more strongly on the QUDs for the narrative than for the car review. This is consistent with both the mean number of different QUDs produced per statement and the related mean entropy values per text type: In the car review, 18.58 distinct QUDs were produced on average and the mean entropy is 3.81. Both interrelated figures are higher than in the narrative, where on average 14.6 QUDs were produced and the mean entropy is 3.22.

We also explored two potential explanations for this difference: (i) a difference in sentence complexity between text types and (ii) the position of the assertion in the text.

Sentence complexity might affect the variation between QUDs because the more words a sentence contains, the more different focus-background structures (related to different QUDs) it can have. For this reason, we approximated the complexity of an assertion with the number of words it contains, which is higher for the car review (mean = 17, sd = 7.71) than for the narrative (mean = 7.25, sd = 5.5). However, as Figure 2 (**A**) suggests, these measures are not correlated in any of the two texts ($r_{narrative} = 0.1$, $p > 6$; $r_{review} = 0.18$, $p > 0.4$). This suggests that the entropy is influenced by other factors than the pure complexity (at least if measured as the word count per assertion).

The entropy in the distribution over QUDs might be affected by the position of an assertion in a text in two ways. In principle, it could (i) decrease as the texts proceeds since the discourse structure becomes clearer and the QUDs on average more predictable or (ii) increase, because more sentences, which might each give rise to further QUDs, have been processed. Figure 2 (**B**) illustrates the correlation between an assertion's entropy and the position of this assertion in the text for both texts. While for the narrative the correlation is again negligible ($r = -0.02$, $p > 0.9$), it appears to be moderate for the car review ($r = 0.43$, $p > 0.06$). The positive direction of this correlation hints at the entropy increasing in the course of the text, which is in line with the assumption that a content-wise increase of complexity leads to greater uncertainty about the current QUD. Of course these are only preliminary observations that need to be validated on larger proportions of more texts.

7 Summary

Our contribution to the QUD-Anno Challenge is a crowdsourced but manually annotated data set for the beginning of the narrative and the car review that allows for calculating probability distributions over possible QUDs. Our data set shows that while there is often one preferred, i.e. most probable, QUD per assertion, there still exits considerable variation among the QUDs that naive annotators assume. Our contribution allows us to reveal and quantify this variation through probabilities and entropy as measures and at the level of individual assertions, between multiple assertions of a text and between the two texts examined. In this way, we believe that our quantitative approach can be a useful complement to the more fine-grained qualitative expert annotations. Our data set allows for the investigation of a series of further research questions, like e.g.:

- What is the relation between the QUDs that experts assign to an assertion based on theoretical assumptions and QUDs that are produced by naive participants?
- To what extent do the experts' gold standard QUDs match the most frequently produced QUD in our data set?
- What could be theory-relevant properties of an assertion that determine the entropy within the distribution over its possible QUDs?

Given an appropriate information-structural annotation, we could also investigate:

- Which role do information-structural categories such as focus, topic and (non-)at-issueness play for the QUDs produced by our participants ?
- Is there a match between the (assumed) focus-background structure of the assertion and the produced question?
- Did participants pay attention to topic chains established in a text?
- Are there questions that address information that is considered to be not at issue?

References

De Kuthy, K., Reiter, N., & Riester, A. (2018). QUD-based annotation of discourse structure and information structure: Tool and evaluation. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Karttunen, L. (1977). Syntax and Semantics of Questions. *Linguistics and Philosophy*, *1*(1), 3–44. https://doi.org/10.1007/BF00351935

Morgan, J. (1973). Sentence fragments and the notion 'sentence'. In B. B. Kachru, R. Lees, Y. Malkiel, A. Pietrangeli, & S. Saporta (Eds.), *Issues in linguistics. Papers in honor of Henry and Renée Kahane* (pp. 719–751). University of Illionois Press.

Poppels, T., & Kehler, A. (to appear). Ellipsis and the QUD: Sluicing with Nominal Antecedents. In A. Konietzko & S. Winkler (Eds.), *Information Structure and Discourse in Generative Grammar: Mechanisms and Processes*. De Gruyter Mouton.

Reich, I., Lemke, R., & Schäfer, L. (to appear). Questions under discussion, salience and the acceptability of fragments. In A. Konietzko & S. Winkler (Eds.), *Information Structure and Discourse in Generative Grammar: Mechanisms and Processes*. De Gruyter Mouton.

Riester, A. (2019). Constructing QUD trees. In M. Zimmermann, K. von Heusinger, & E. Onea (Eds.), *Questions in Discourse. Volume 2: Pragmatics* (pp. 164–193). Brill. https://doi.org/10.1163/9789004378322_007

Riester, A., Brunetti, L., & De Kuthy, K. (2018). Annotation guidelines for Questions under Discussion and information structure. In E. Adamou, K. Haude, & M. Vanhove (Eds.), *Information Structure in Lesser-described Languages: Studies in prosody and syntax* (pp. 403–444). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.199.14rie

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*, 1–69. https://doi.org/10.3765/sp.5.6

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

van Kuppevelt, J. (1995). Discourse structure, topicality and questioning. *Journal of Linguistics*, *31*(1), 109–147. https://doi.org/10.1017/S002222670000058X

von Stutterheim, C., & Klein, W. (1989). Referential movement in descriptive and narrative discourse. In R. Dietrich & C. F. Graumann (Eds.), *Language processing in social context* (pp. 39–76). North Holland.

Westera, M., Mayol, L., & Rohde, H. (2020). TED-Q: TED talks and the questions they evoke. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1118–1127.

Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832