TED-Q: TED talks and the questions they evoke

Westera, M.; Mayol, L.; Rohde, H.; Calzolari, N.; Béchet, F.; Blache, P.; ... ; Piperidis S.

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1118–1127
Marseille, 11–16 May 2020

TED-Q: TED Talks and the Questions they Evoke

Matthijs Westera, Laia Mayol, Hannah Rohde
Universitat Pompeu Fabra (×2), University of Edinburgh
Barcelona (Spain), Edinburgh (Scotland)
{matthijs.westera, laia.mayol}@upf.edu, hannah.rohde@ed.ac.uk

Abstract

We present a new dataset of TED-talks annotated with the questions they evoke and, where available, the answers to these questions. Evoked questions represent a hitherto mostly unexplored type of linguistic data, which promises to open up important new lines of research, especially related to the Question Under Discussion (QUD)-based approach to discourse structure. In this paper we introduce the method and open the first installment of our data to the public. We summarize and explore the current dataset, illustrate its potential by providing new evidence for the relation between predictability and implicitness – capitalizing on the already existing PDTB-style annotations for the texts we use – and outline its potential for future research. The dataset should be of interest, at its current scale, to researchers on formal and experimental pragmatics, discourse coherence, information structure, discourse expectations and processing. Our data-gathering procedure is designed to scale up, relying on crowdsourcing by non-expert annotators, with its utility for Natural Language Processing in mind (e.g., dialogue systems, conversational question answering).

Keywords: Discourse structure, discourse relation, evoked question, question under discussion, TED-talks, crowdsourcing, implicit connective

1. Introduction

Discourse structure is relevant for a variety of semantic and pragmatic phenomena and is increasingly important for a number of language technologies. It is integrated into theoretical and psycholinguistic models of a range of context-driven effects (Cummins and Rohde, 2015), including those in coreference (Kehler and Rohde, 2013; Polanyi, 1988), presupposition (Kim et al., 2015), implicature (Beaver and Clark, 2008), discourse particles and cue phrases (Hirschberg and Litman, 1993), among others. Within computational systems, multiple domains rely on semantic resources to support the derivation of meaning in text processing and to produce natural sounding language in generation tasks. Discourse structure informs applications such as anaphora resolution (Voita et al., 2018), argument mining (Hewett et al., 2019), machine translation (Xiong et al., 2019), and text simplification (Siddharthan, 2003).

One way of articulating the structure of a text is to identify the questions and subquestions that are raised and answered by subsequent spans of text. Models of Questions Under Discussion (QUDs) posit underlying structures that are built around a sequence of discourse moves consisting of questions and their answers (Carlson, 1983; Ginzburg, 1994; Ginzburg and Sag, 2000; van Kuppevelt, 1995; Larsson, 1998; Roberts, 1996). These questions and answers can be understood in terms of their use in moving a discourse forward to achieve communicative goals and subgoals. QUDs influence both the surface form of the answer and the meaning derived from that answer. But not all QUDs are explicit, in fact most are not, particularly in natural discourse. Recovering implicit QUDs is therefore key for understanding the underlying discourse structure of a text and for the use of such structure in modeling other phenomena.

The current work offers a new methodology for the elicitation of human judgments on QUD predictability with the aim of giving researchers access to a large-scale window on discourse structure. More precisely, we probe what questions a discourse evokes and subsequently which of those are taken up as the discourse proceeds. The primary contributions of this work are the scalability of the methodology and the augmentation of an existing discourse-structure-annotated resource TED-MDB (Multi-lingual Discourse Bank) (Zeyrek et al., 2018) with a new annotation layer (which we term TED-Q), released here as a preliminary dataset for the public. We illustrate the potential of this new resource by exploiting the double annotation layer via a novel empirical demonstration of the oft-posited link between predictability and reduction (Levy and Jaeger, 2007; Aylett and Turk, 2004): We identify QUD predictability with the degree to which our annotators' questions ended up being answered, and establish robust patterns of reduction (lower rates of explicit marking of discourse relations) at text positions where the QUD was more predictable.

Our TED-Q dataset offers a new type of cognitive/linguistic data for language technologies, one with the potential to open up and connect several lines of research. It should be of interest, at its current scale, to researchers on formal and experimental pragmatics, discourse coherence, information structure, discourse expectations and processing, and question-answer systems. Moreover, our data-gathering procedure is designed to scale up, with its utility for NLP in mind. We release the TED-Q dataset, annotation interfaces and analysis scripts on `https://github.com/amore-upf/ted-q`.

2. Background

Questions Under Discussion (QUDs) offer an open-ended discourse-structuring device, with no set inventory of possible questions or sub-questions. This means that annotating discourse structure using QUDs can be (in part) a matter of entering free-form questions at places in the discourse (De Kuthy et al., 2018). In this respect QUD-based models differ from many theories of discourse structure, particu-

larly those that rely on a finite inventory of possible discourse relations. These relation-based approaches to discourse structure and coherence have a long history, with a variety of different posited inventories of possible relations (see Knott (1996); for corpus-based comparisons of different annotation schemes, see Wolf and Gibson (2005) and Sanders et al. (2018)). These inventories can be large and sophisticated, making it hard for non-expert annotators to choose the right discourse relation – though the Penn Discourse TreeBank (PDTB) annotation scheme (Prasad et al., 2019) partially overcomes this by associating relations with linguistic connectives such as "because" and "however". By contrast, entering a free-form question that connects two pieces of discourse can be a more natural task, as noted also in Anthonio et al. (2020).

Theories of discourse structure often acknowledge both a local structure, relating one utterance and the next, and an overarching structure, relating longer stretches of discourse to each other and/or to overarching goals. QUD-based theories typically assume that QUDs are organized in a discourse tree structure, with a super-question at the top and sub-questions towards the bottom (Roberts, 1996). Some relation-based theories posit discourse relations both between individual discourse segments and between larger chunks of multiple segments joined together (Asher and Lascarides, 2003; Hobbs, 1979; Kehler, 2002; Mann and Thompson, 1988), likewise giving rise to a hierarchical structure. The PDTB (Prasad et al., 2019) approach is instead restricted to more local relations, by considering explicit or inferable connectives between clauses, remaining agnostic about any overarching discourse structure.

In this work, we present an annotation task for local discourse structure expectations based on the QUD-approach. More precisely, we present annotators with local pieces of discourse and ask them which question a passage *evokes* (cf. 'potential questions' of Onea (2016)). Subsequently we show them how the discourse continues and ask them whether their question has been answered. This local, incremental, two-step annotation process is suitable for non-expert annotators, as the individual steps are small, intuitive tasks. This lets us avoid the well-known pitfalls of reliance on expert annotators concerning scalability, cost and theoretical bias (see similar arguments for the connective-insertion tasks used by Yung et al. (2019; Rohde et al. (2018)). It makes our dataset of evoked questions comparable in this regard to, e.g., large-scale word similarity benchmarks, which are compiled not from a handful of trained experts but from a large number of theory-neutral individuals who are asked to make local, intuitive judgments.

Another core motivation for this incremental, two-step process is that it gives us a window on QUDs and QUD predictability. If a discourse at a certain point reliably *evokes* a certain question, and subsequently proceeds to *answer* that question, then that question is very likely to be the QUD at that point. To illustrate:

(1) I noticed the teacher scolded the quiet student after class because *the student slept through the lecture*.

If you read only the first clause, the underlined parts will likely evoke a question about WHY the described situation has arisen. This question then ends up being answered by the second clause as you read on (in italics), making it a plausible QUD for that clause. The degree to which evoked question end up being answered as the discourse unfolds is a measure of the predictability of QUDs.

Prior work on discourse structure annotation does not take this incremental, forward-looking approach, wherein subsequent discourse is hidden until a question is posed. Instead, QUD recovery has been treated as a predominantly backward-looking process: each utterance is analysed to establish what *prior* question it answers relative to the preceding context (Anthonio et al., 2020), or even with respect to content in the entire preceding and subsequent discourse (De Kuthy et al., 2018; Riester, 2019), rather than which new question it evokes (Onea, 2016). In our case annotators have less information to work with, as the continuation of the discourse is hidden until they pose a question. This inevitably results in less complete QUD recovery, but it does make our annotation task more natural (quite like engaging in ordinary dialogue), and furthermore it uniquely provides a window on QUD predictability in the way described above, on which we will capitalize in the present paper.

Given our research aim of using evoked questions as a window on QUD predictability and discourse structure more generally, we chose to annotate a corpus that comes with existing discourse structure annotations: TED-MDB (Multi-lingual Discourse Bank), a set of TED talks with PDTB-style discourse relation annotations (Zeyrek et al., 2018). Crucial for our aim is that discourse relations and QUDs, although belonging to different frameworks, are closely related (Kehler and Rohde, 2017). For instance, in (1), the causal relation (signaled in this case with the explicit marker *because*) corresponds to the 'Why?' question raised by the first clause and answered by the second. Another advantage of the TED-MDB corpus is that it consists of reasonably naturalistic (though rehearsed) spoken language, which is important given the growing emphasis in the field on naturalistic text. TED talks offer a middle ground between written genres in newspaper or academic texts and the fully-open ended nature of unscripted dialogue.[1] This affords us the opportunity to test our new method on the kind of data that will help inform generative, open-ended models of QUD prediction.

Our evoked questions stand to inform the semantic and pragmatic theories that rely on QUD-based discourse structure (e.g., the status of a QUD-dependent presupposition may vary with the predictability of that QUD). In addition, we are interested in QUD predictability itself as a domain of inquiry for testing models of linguistic redundancy and efficiency. As noted earlier, predictability is associated with reduction, such that more predictable linguistic elements are candidates for reduction or omission

[1] We piloted our methodology with another, more spontaneous, unscripted spoken corpus, DISCO-SPICE (Rehbein et al., 2016), but it posed a number of challenges that are typical of fully unscripted discourse.

during language production; this pattern is often referred to as, among other names, the Uniform Information Density Hypothesis (Levy and Jaeger (2007); see also Aylett and Turk (2004)). Evidence for this generalization has been found at the level of sound (Turnbull, 2018), words (Gahl, 2008), syntax (Frank and Jaeger, 2008), and discourse relations (Asr and Demberg, 2012). QUDs represent an understudied linguistic representation over which language users may compute predictability. Their surface realization via explicit discourse markers (e.g., *because* in (1)) is crucially optional in many cases, raising the possibility that these optional markers will be omitted at higher rates on utterances for which the predictability of the question being addressed is higher. Our new methodology makes it possible to generate estimates of QUD predictability to test this hypothesis.

3. Method

As our starting dataset we use TED-Multilingual Discourse Bank (MDB) (Zeyrek et al., 2018). It consists of transcripts of six scripted presentations from the TED Talks franchise, in multiple languages, but we will use only the English portion (6975 words total). Zeyrek et al. annotated these transcripts with discourse relations, in the style of PDTB (Prasad et al., 2019), and we will rely on this for some analysis in section 5.. Earlier pilots we conducted relied on unscripted spoken dialogues from the DISCO-SPICE corpus (Rehbein et al., 2016), but these transcripts were too hard to follow for our participants. Relying on the scripted presentations of TED-MDB avoided this problem while still remaining in the realm of reasonably naturalistic spoken text.

Our contribution is to extend this existing dataset with elicited questions. Our procedure consists of two phases: the *elicitation phase* where we ask people to read a snippet of text and enter a question it evokes, then read on and indicate whether the question gets answered and how, and a *comparison phase* where we ask people to indicate which of the elicited questions are semantically/pragmatically equivalent, or more generally how related they are. The second phase is necessary because in the first phase we elicit questions in free-form, and what counts semantically/pragmatically as 'the same question' can be operationalized in many different ways. We will describe each phase in turn.

Elicitation phase For the elicitation phase, texts were cut up into sentences (using NLTK's sentence tokenizer), and long sentences only (> 150 words) were further cut up at commas, colons or semicolons by a simple script.[2] For convenience we will refer to the resulting pieces of text as sentences. Our aim was to fully cover the TED-MDB texts with evoked questions, by eliciting evoked questions after every sentence. We decided to present excerpts of these texts instead of full texts, because we wanted our approach to be able to scale up to (much) longer texts in principle,

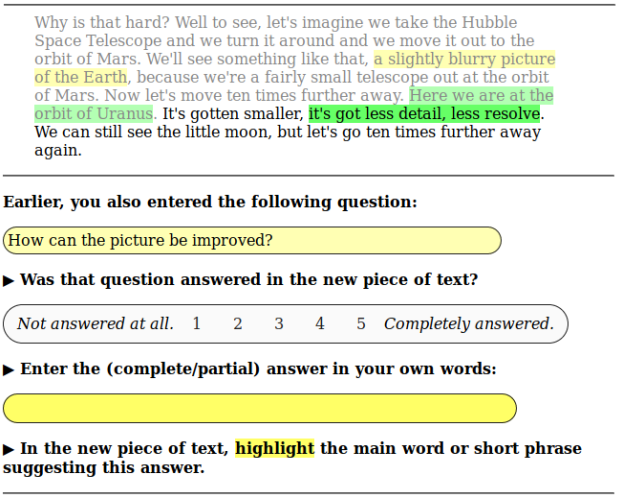

Figure 1: A view of our elicitation tool, here asking whether a previously entered question has been answered yet.

and in order to keep annotators fresh. We presented each participant with up to 6 excerpts from different source texts (more would have made the annotation task too long), each excerpt comprising up to 18 sentences (a trade-off between having enough context and keeping annotators fresh). Each excerpt was incrementally revealed, with a *probe point* every 2 sentences. To still get full coverage of the texts we alternated the locations of probe points between participants. In this way we covered the 6975 words of TED-MDB with a total of 460 probe points.

At each probe point participants were asked to enter a question evoked by the text up to that point, and, for previously unanswered questions evoked at the previous two probe points, they were asked whether the question had been answered yet by choosing a rating on a 5-point scale from 1 'completely unanswered' to 5 'completely answered' (henceforth ANSWERED). We limited the number of revisited questions to 2 in order to avoid breaking the flow of discourse too much and to prevent the task from becoming too tedious, although this may mean that we will miss some answers. (However, in a pilot study we found that questions that weren't answered after the first probe point wouldn't be answered at the next two probe points either.) The formulation asking for evoked questions was: "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!)". The screen for indicating answers is shown in figure 1.

The decision to present only excerpts, and to check question answeredness only for two subsequent chunks, make scalable annotation by non-experts feasible. However, this biases our approach towards questions that reflect only 'local' discourse structure. This restriction must be kept in mind, but note that our approach shares this locality for instance with the discourse relations approach, and accordingly with the existing annotations of TED-MDB on which we will rely further below. For a detailed overview of our elicitation phase and more reflection on design decisions such as these, we refer to an earlier report (Westera and Rohde, 2019).

[2] Neither the original sentences nor the pieces into which we cut longer sentences necessarily correspond to what are sometimes called discourse segments, though often they do. On some occasions this makes our coverage of the existing discourse relation annotations lower than it could have been.

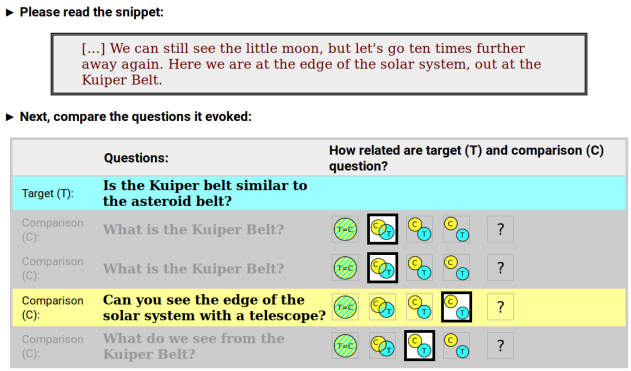

Figure 2: A view of our comparison tool; participants had to click to reveal the questions; the yellow highlighting follows the cursor, helping to focus each comparison.

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 did this by dragging a selection in the newest two sentences of the excerpt, and could highlight at most 10 words. The motivation behind this word limit was that it would force annotators to be selective, thus making their highlights more informative (we want only the most important words, even if without context these would not suffice to evoke the question or provide the answer in full). Highlights for different questions were given different colors, and highlights for answers were given the same color as the question they were answers to.

We set up this task in Ibex (Internet-based experiments, `https://github.com/addrummond/ibex/`), hosted on IbexFarm (`http://spellout.net/ibexfarm/`), and recruited 111 participants from Amazon Mechanical Turk (`https://www.mturk.com/`).[3] Each participant could do the task once. We estimated that the task would take about 50 minutes, and offered a monetary compensation of $8.50. 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. For an overview of these basic numbers (as well as the resulting data, discussed in the next section) see Table 1.

Comparison phase The goal of the comparison phase, recall, was to establish a notion of inter-annotator agreement on the (free-form) questions we elicited, by gathering judgments of question relatedness/equivalence. For this, we set up a task in the Mechanical Turk interface directly. A screenshot is shown in figure 2. We published tasks of 10 snippets of around 2 sentences, each followed by an exhaustive list of the questions we elicited at that point. In each task one of these questions was designated the 'target question', the others 'comparison questions', and participants were asked to compare each comparison question to the target question. Questions were rotated through the 'target question' position, so for every pair of questions we would get the same number of comparisons in either order. For each comparison our participants were instructed to select one of the following options (the Venn-diagram-like icons from left to right in the image):

- *Equivalence*: Target and Comparison question are asking for the same information, though they may use very different words to do so.
- *Overlap*: Target and Comparison question are slightly different, but they overlap.
- *Related*: Target and Comparison question are quite different, no overlap but still closely related.
- *Unrelated*: Target and Comparison question are very different; they are not closely related.
- *Unclear*: Target and/or Comparison question are unclear.

In addition to these descriptions, we instructed participants that what we were after is "what kind of information the questions are asking for, not how they are asking it", with the advice to look beyond superficial appearance, to interpret the questions in the context of the text snippet, and that if two questions invite the same kinds of answers, they count as the same kind of question.

We estimated that each task would take around 4 minutes and offered a reward of $0.90. We limited participants to doing at most 20 tasks per person (each task consisting of 10 snippets) to ensure diversity. We ended up recruiting 163 workers. For these basic numbers (as well as numbers of the resulting data, discussed next), see again Table 1.

4. The resulting dataset: TED-Q

Results of elicitation phase Our elicitation phase resulted in 2412 evoked questions, 1107 annotations that a previously elicited question was at least partially answered by a given subsequent chunk (ANSWERED $\geq$ 3 on the scale from 1 'completely unanswered' to 5 'completely answered'), and 2562 annotations that a previously elicited question was not answered by a given subsequent chunk (ANSWERED $<$ 3). For the basic numbers see table 1. Both questions and answers contain both the free-form question/answer as entered by the participant, and the words in the chunk which the participant highlighted as primarily evoking the question/providing the answer, respectively. On average participants highlighted 5.2 words for questions and 5.6 words for answers (standard deviation for both is 2.5).

Recall that any question evoked by a chunk, according to a worker, was presented to that same worker in up to two subsequent chunks, to see whether it has been answered. As the ANSWERED rating of a question we take the highest ANSWERED rating achieved by its two subsequent chunks. Averaged across all evoked questions this ANSWERED rating is 2.50 (standard deviation 1.51), so questions tend towards remaining unanswered. Still, almost half of the questions

[3] One further participant was excluded for only entering their questions as a single, all-caps word; the numbers reported concern the remaining data (N=111).

Elicitation phase:		Comparison phase:	
texts:	6	question pairs:	4516
words:	6975	participants/pair:	6
probe points:	460	participants:	163
participants/probe:	5+	judgments:	30412
participants:	111	RELATED mean:	1.21
questions:	2412	RELATED std:	0.79
answers:	1107	Agreement (AC_2):	.46
ANSWERED mean:	2.50		
ANSWERED std:	1.51		

Table 1: Basic numbers of the TED-Q dataset.

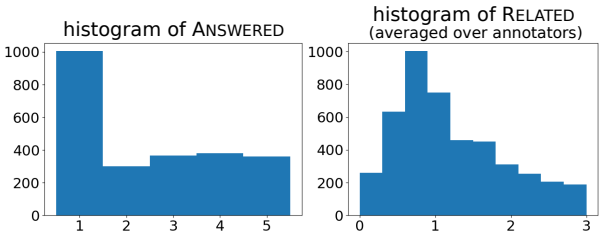

Figure 3: Distributions of ANSWERED judgments (elicitation phase) and RELATED scores (comparison phase, averaged over annotators).

(1107) are at least partially answered, with 367 completely answered; see the first histogram in figure 3. We think that this proportion is quite high, given the 'locality' of our elicitation method – recall that unanswered evoked questions were revisited at most twice and then dropped. It suggests that participants ask questions that anticipate speakers' upcoming discourse moves, although as expected there is also considerable indeterminacy.[4]

We also looked at the distribution of elicited 'question types', which we defined essentially by the first word of the question, though taking some multi-word expressions into account as well (e.g., we analyze "how come" as the same type as "why", not as "how"). The distribution of question types is shown in figure 4. *What*-questions were the most frequent, likely due to the flexibility of this wh-word. Auxiliary-initial polar questions were next, followed by *how/why*-questions (setting aside the 'other' class, which is in need of further analysis; it contains for instance declarative echo questions). *Where/who*-questions are often meta/clarification questions (e.g., Who are they talking about? Where are they?). Breakdown of ANSWERED by question type suggest that the latter are also the least answered – likely reflecting that our participants' meta/clarification questions were not as at-issue for the original speaker – together with *when*-questions. *Why/what* questions were the most answered (after 'other'), suggesting more reliable QUD anticipation. This is shown in figure 5. Most differences in the plot involving one or two of the larger classes are significant (t-test, $p < .05$), but among

[4] We agree with an anonymous reviewer that it could be useful to have a portion of the data annotated by experts, or ourselves, for comparison, but so far we have not done this.

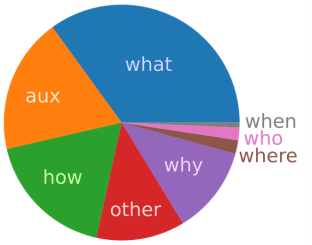

Figure 4: Distribution of question types based on initial word (and some multi-word expressions).

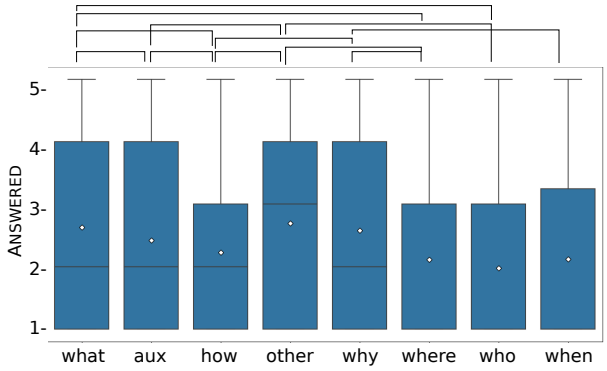

Figure 5: ANSWERED score per question type; boxes show the middle quartiles with a horizontal line for the median (ANSWERED median is 1 for 'where', 'when', 'who'), a white marker for the mean. Braces mark significant differences of ANSWERED (t-test, $p < 0.05$).

the smaller classes (*where*, *who*, *when*) we lack statistical power; the braces on top indicate significant differences.

Results of comparison phase For the subsequent comparison phase, we took, for every probe point, all pairs of evoked questions that were entered at that point, resulting in 4516 question pairs (453 probe points times (mostly) $(5 * 4)/2$ pairs of questions per probe point). These were given to a new batch of participants for annotating question relatedness in six-fold (each pair three times in either order), resulting in a total of 30412 annotations by 163 participants. Average RELATED rating is 1.21 (average standard deviation per question pair is 0.79) on a scale we represent numerically from 0 to 3 (0 = not closely related; 3 = equivalent), which means that on average questions were judged as 'closely related but no overlap'; see the second histogram in figure 3. Inter-annotator agreement is .46 using the metric AC_2 with quadratic weights (Gwet (2014); we used the R package `irrCAC`), which is more paradox-resistant than for instance Cohen's κ or Scott's π, and which can handle different annotators covering different portions of the data. This represents 'moderate' agreement according to Landis and Koch (1977), which for the present task (and after manual inspection of some examples) we think is acceptable, given its subjectivity (Craggs and Wood, 2005), e.g., questions often permit multiple interpretations. Here is an example of a probe point with high RELATED rating among the evoked questions, as well as high ANSWERED rating (the age question

		Spearman:
RELATED	GLEU	0.47
RELATED	SIF	0.32
RELATED	SAME-WH	0.24
ANSWERED	RELATED	0.17
ANSWERED	GLEU	0.096
ANSWERED	SIF	0.078
ANSWERED	SAME-WH	0.045

Table 2: Correlations between RELATED and ANSWERED, and various computational notions of question similarity.

makes sense only in the broader discourse, omitted here):

(2) [...] one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses. `ted_mdb_1971`
Why wouldn't they use their prostheses? / Why do they not use their prostheses? / Did they do something to help amputees? / How old are you now? / Why didn't amputees use their prostheses?
The reason, I would come to find out, was that their prosthetic sockets were painful [...]

We tested whether our human RELATED scores could have been replaced by an automatic method. The top portion of Table 2 shows Spearman correlation coefficients between RELATED and three automatic measures: GLEU, which is indicative of surface-structure similarity (Wu et al., 2016); SIF, which represents distributional semantic similarity, computed as the cosine between the evoked questions' Smooth Inverse Frequency embeddings (Arora et al., 2017), which are high-dimensional, distributional semantic vector representations; and SAME-WH, which is a binary attribute representing simply whether questions belong in the same class according to our coarse classification (i.e., the classes shown in figure 4). As expected all of these automatic measures correlate with RELATED, though no correlation is particularly strong. For the surface-oriented scores GLEU and SAME-WH this is because what is semantically/pragmatically the same question can be asked in many different ways; here is an example from our dataset with high RELATEDness which the automatic scores miss:

(1) [...] In Navajo culture, some craftsmen and women would deliberately put an imperfection in textiles and ceramics. `ted_mdb_1978`
*What does Navajo culture have to do with the matter at hand? / How does that apply here? (*RELATED*:* 2.50*;* GLEU*: 0.04;* SIF*: 0.46;* SAME-WH*: 0)*

We hope that our RELATED scores will offer a useful new human benchmark for evaluating sentence similarity and sentence embedding methods from Computational Linguistics. For one, questions are underrepresented in existing datasets, which tend to focus on assertions (e.g., inference benchmarks (Bowman et al., 2015)). An important feature of our dataset in this regard is that the relatedness judgments are *contextualized* (e.g., McDonald and Ramscar (2001)): the evoked questions often contain anaphoric elements such as pronouns and ellipsis, relying for their interpretation on the snippet that evoked them (recall that those snippets were given also in the comparison phase of our crowdsourcing process). Such context-dependence is well-known to yield additional challenges for computational methods. But at present we will not further explore this possible use of our TED-Q dataset.

Recall our motivating assumption from section 1., that a question that is both reliably evoked by the preceding discourse and answered by its continuation, is likely a Question Under Discussion at that point. The foregoing results lend us two indicators of the predictability of a Question Under Discussion: high RELATED ratings indicate that a certain kind of question is reliably evoked by a discourse, and high ANSWERED ratings indicate whether those questions were answered. We expect to see a correlation between RELATED and ANSWERED, where the strength of this correlation is a measure of how predictable Questions Under Discussion are: if reliably evoked questions tend to be answered most of the time (and non-reliably evoked questions tend not to), that means the Question Under Discussion is generally predictable from the prior discourse. Indeed, we find a weak but significant Spearman correlation between RELATED and ANSWERED (correlation coefficient 0.17, $p = 3\text{e-}16$). See the lower part of Table 2, also for a comparison to correlations of ANSWERED with surface form similarity (GLEU), distributional semantic similarity (SIF) and sameness of *wh*-word. These correlations further affirm that the comparison phase of our crowdsource method has added value: the human relatedness judgments give us something different from the automatic measures.

5. Using TED-Q for quantifying anticipation of TED-MDB's discourse relations

The main reason we selected our source texts from the TED-MDB dataset is that they have already been annotated with discourse structure (Zeyrek et al., 2018). Our contribution of TED-Q therefore enables us to investigate the relationship between discourse structure and the evoked questions we elicited, a relationship which should be close given the close connection between evoked questions and potential/actual Questions Under Discussion (QUD) as used in the QUD-based approach to discourse structure. TED-MDB annotates discourse structure by identifying discourse relations between adjacent clauses, using the taxonomy of the Penn Discourse Treebank (PDBT) (Prasad et al., 2019). Combining TED-MDB with TED-Q gives us decent dual coverage: 84% of the questions we elicited were produced at a point where TED-MDB has an annotation for the relation holding between the fragment immediately preceding the question and the fragment immediately following it; conversely, 62% of the discourse relation annotations correspond to our probe points (since we wanted to incrementally present only complete(ish) sentences to our participants, we miss occurrences primarily of clause-internal connectives such as "but" and "and").

The PDTB-style annotation used in TED-MDB has several levels. At the most general level, the type of relation

holding between each pair of adjacent arguments is annotated using one of the following categories: Explicit (if there is a connective expresses the discourse relation), AltLex (if an expression other than a connective expressing the discourse relation), Implicit (there is a discourse relation but it is not lexically expressed), EntRel (there is not a discourse relation, but the arguments are related by mentioning the same entity) and NoRel (there is no relationship between the sentences). If there is a discourse relation (i.e., Explicit, Implicit or AltLex), it is further categorized as either Temporal, Contingency (one argument provides the reason, explanation or justification of the other), Comparison (the two arguments highlight their differences or similarities), Expansion (one argument elaborates on the other) or Hypophora (Question-Answer pairs). Each of these categories is subdivided into several subtypes of discourse relations, some of which are further subcategorized, e.g., Temporal.Asynchronous.Precedence or Contingency.Cause.Result.

A natural thing to look for is a correlation between the type of discourse relation holding between two sentences and the type of evoked questions we elicited at that point (i.e, directly after the first sentence, before the second sentence was shown). The best candidate to do so are *why*-questions, since they are strongly linked to a particular discourse relation (i.e. causality), as opposed to other *wh*-words which may have many different uses (*what* and *how*) or are not clearly associated with a discourse relation (*when* and *where*). A clear correlation emerges between *why*-questions and causal relations (Cause, Cause+Belief and Purpose); while the overall proportion of *why*-questions is 12%, this goes up to 19% at points where the relation is causal (significantly so: χ^2 (7, N = 1580) = 20.58, p $< .01$). Thus, even with a simple classification of question types (initial word), we find some evidence for the expected correlation between the kinds of questions evoked at a given point and the upcoming discourse relation.[5]

Pending a more precise classification of question types, there are more general patterns to observe: For instance, questions that were evoked at a point annotated as NoRel exhibited significantly lower ANSWERED and RELATED scores than questions evoked when there was a relation: they were answered less (t(2219)= 4.71, $p < .0001$) and were less related to each other (t(2219)= 4.23, $p < .0001$). This suggests that it is harder to anticipate the QUD at those points in the discourse where the current sentence and the next are not directly related to each other.

In the remainder of this section we will use the TED-Q/TED-MDB alignment to investigate an influential linguistic hypothesis: the Uniform Information Density (UID) Hypothesis (Frank and Jaeger, 2008). It states that the rate of information exchange tends to be kept constant throughout an utterance or discourse. Asr and Demberg (2012) note that the UID Hypothesis entails that discourse relations that are more predictable will tend to be more implicit. To test this hypothesis, Asr and Demberg needed to rely on prior assumptions about which relations are on the whole more predictable (the Causality-by-default Hypothesis and the Continuity Hypothesis, discussed separately further below). By contrast, TED-Q uniquely enables us to quantify the predictability of discourse relations in a data-driven way, namely, in terms of the ANSWERED scores of evoked questions; moreover, this notion of predictability is context-dependent: a given type of relation may be predictable in some contexts and unpredictable in another.

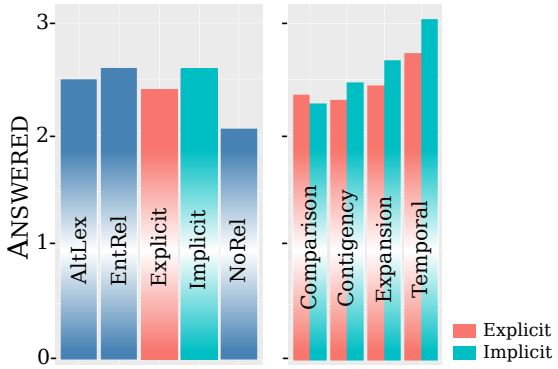

Figure 6: ANSWERED scores across types of relations and across types of implicit and explicit discourse relations

Using TED-Q we find direct support for Asr and Demberg's prediction: Questions produced where there was an Explicit relation indeed end up being answered significantly less (signifying unpredictability) than questions produced where the relation was Implicit (t(1570)=2.39, p=.016).[6] See figure 6 for the mean ANSWERED score of questions evoked at different types of relations (left), and a closer look comparing Implicit and Explicit discourse relations of each type (right). Thus, TED-Q can be used to quantify predictability of discourse structure, in a data-driven way, without relying on the two assumptions about predictability used in Asr and Demberg (2012), namely, the Causality-by-default hypothesis and the Continuity Hypothesis. This is welcome, because evidence for these in TED-MDB/TED-Q is weak, as we show in the remainder of this section.

The Causality-by-default Hypothesis (Sanders, 2005) postulates a general preference for causal relations. In support of this, Asr and Demberg (2012) report that the Cause relation is the most frequent implicit relation in PDTB, and also the (frequent) relation that has the highest implicitness (65% of Cause relations are implicit). In TED-MDB this picture is less clear: Although Cause (including Belief/SpeechAct variants) is the most frequent

[5] We are planning a third round of annotations aimed at categorizing our evoked questions more semantically/pragmatically, using a taxonomy resembling the PDTB inventory of discourse relations, so that more correlations can be examined.

[6] The effect is understandably small, because discourse anticipation is hard and many evoked questions inevitably remain unanswered. By concentrating on probe points with high RELATEDness, i.e., where people agreed about the evoked question, we see the difference between Implicit and Explicit increase, e.g., for RELATED > 1.5 (3rd quartile, 542 questions), mean ANSWERED for Implicit increases from 2.59 to 3.13, while for Explicit it stays roughly the same (2.41 and 2.48, respectively).

implicit relation in TED-MDB, this is not by as large a margin (22%, followed at 21% by Conjunction and Level-of-Detail); and although the implicitness of Cause relations in TED-MDB (50%, vs. 65% in PDTB) is still higher than average, it is not the highest among the frequent relations. As for TED-Q, the Causality-by-default Hypothesis leads one to expect that causal questions get asked and/or answered more, but neither is decisively the case. For one, although *why*-questions (in which we included variants "how come", "what for", "for what reason") are indeed among the most answered (Figure 5), their ANSWERED score is slightly (non-significantly) lower than "what" and "other", and not significantly higher than polar questions ("aux") either. Moreover, whereas causal relations are the most frequent implicit relation, *why*-questions (including "how come", etc.) are with 12% only the fourth most frequent question type, after *what*-questions, polar questions and *how*-questions (see Figure 4). Note that no strong conclusion should be drawn from this, given our coarse classification of questions and given that the more frequent *what*-questions and polar questions are both very heterogeneous classes.

The Continuity Hypothesis (Segal et al., 1991; Murray, 1997) postulates a preference for (hence greater predictability of) continuous relations and temporal relations that are ordered linearly. In support of this, Asr and Demberg (2012) found that in PDTB continuous relations (Cause, Instantiation and Level-of-detail) are more often implicit than discontinuous ones, and relations that have both 'forward' and 'backward' versions (Cause, Concession and Asynchronous) are more implicit in their forward version. But although the relation counts in TED-MDB reveal mostly the same pattern (omitting details for reasons of space), the ANSWERED scores in TED-Q do not. The Continuity Hypothesis predicts that questions evoked prior to a continuous or forward relation should have a higher ANSWERED score, but this is not the case: we find no significant effect of continuity ($t(1570)= 1.43$, $p = .15$), nor of forward/backward ($t(257)= 0.81$, $p = .41$, for Cause; we have insufficient data for Concession and Asynchronous).

Summing up, by quantifying the predictability of a discourse relation as the rate by which evoked questions in TED-Q were answered we were able to confirm the UID Hypothesis, i.e., that discourse relations are more often implicit when they are predictable, though with only weak, partial support for its two sub-hypotheses used in Asr and Demberg (2012). This might reflect some inherent difference between the ways in which evoked questions vs. discourse relations reflect discourse structure, or that a context-dependent notion of predictability, such as ANSWERED in TED-Q, is more fine-grained than generalizations such as the Continuity Hypothesis – e.g., continuity may be predictable in some contexts but not in others.

6. Conclusion

While previous work has shown the relevance of Question Under Discussion (QUD)-based approaches for understanding a variety of semantic and pragmatic phenomena, the field has lacked a scalable, non-expert annotation process for QUDs or QUD expectations in naturally occurring discourse. This paper presented a novel methodology for eliciting actual and potential QUDs from non-expert participants. Our annotators were asked simply to enter a question that a short snippet of text evokes for them, and to indicate which words up to that point primarily evoked the question and which words following the question help answer it (if any). The idea behind this method was that questions which are both evoked and subsequently answered are plausible candidates to be the QUD. A separate set of annotators compared the elicited free-form questions, giving us a notion of inter-annotator agreement and an additional way of quantifying QUD predictability. We showed that non-expert annotators indeed pose questions that anticipate speakers' upcoming discourse moves (as measured via the ANSWERED ratings) and which are consistent with those of other annotators (the RELATED ratings).

Altogether this method resulted in the first installment of our TED-Q dataset, which consists of the transcripts of English TED talks annotated with the questions they evoke. This installment contains the six TED-talks of the existing resource TED-MDB, newly annotated with a total of 2412 evoked questions (and their answers and triggers in the text) at 460 probe points, with additional annotations of question relatedness. We release the annotation tools, TED-Q dataset and analysis scripts on `https://github.com/amore-upf/ted-q`.

Because the texts from the TED-MDB corpus have already been annotated with PDTB-style discourse relations, the combination of TED-MDB with TED-Q forms an exciting new resource for the study of discourse structure. We illustrated the potential of this new resource in a number of ways, foremost by offering a new type of evidence for the hypothesis that discourse relations are more often implicit when they are predictable, an instance of the more general relation in natural language between predictability and implicitness. To the extent that our evoked questions represent potential and actual Questions Under Discussion (QUDs), our dataset could be used to shed light furthermore on the relation between these two main approaches to discourse structure, i.e., discourse relations and QUDs.

7. Acknowledgements

We thank the three anonymous reviewers for LREC and also Jacopo Amidei for their helpful commentary. 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) and from the Spanish State Research Agency (AEI) and the European Regional Development Fund (FEDER, UE) (project PGC2018-094029-A-I00). 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.

8. Bibliographical References

Anthonio, T. R., Bhat, I. A., and Roth, M. (2020). wikihowtoimprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'2020)*, Marseille, France, May. European Language Resource Association (ELRA).

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.

Asr, F. T. and Demberg, V. (2012). Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47:31–56.

Beaver, D. I. and Clark, B. Z. (2008). *Sense and Sensitivity: How Focus Determines Meaning*. Wiley-Blackwell, West Sussex, UK.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Carlson, L. (1983). *Dialogue Games: An Approach to Discourse Analysis*. Reidel, Dordrecht.

Craggs, R. and Wood, M. M. (2005). Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3):289–296.

Cummins, C. and Rohde, H. (2015). Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology, Special issue on Context in communication: A cognitive view*, 6:1–11.

De Kuthy, K., Reiter, N., and Riester, A. (2018). 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.

Frank, A. F. and Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Gahl, S. (2008). "time" and "thyme" are not homophones: Word durations in spontaneous speech. *Language*, 84:474–496.

Ginzburg, J. and Sag, I. (2000). *Interrogative Investigations*. CSLI Publications, Stanford.

Ginzburg, J. (1994). An update semantics for dialogue. In *Proceedings of the Tilburg International Workshop on Computational Semantics*.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hewett, F., Rane, R. P., Harlacher, N., and Stede, M. (2019). The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103.

Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19:501– 530.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, 3:67–90.

Kehler, A. and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39:1–37.

Kehler, A. and Rohde, H. (2017). Evaluating an expectation-driven qud model of discourse interpretation. *Discourse Processes*, 54:219–238.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. CSLI Publications, Stanford, CA.

Kim, C. S., Gunlogson, C., Tanenhaus, M. K., and Runner, J. T. (2015). Context-driven expectations about focus alternatives. *Cognition*, 139:28–49.

Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, University of Edinburgh.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, page 849–856.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

McDonald, S. and Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.

Murray, J. D. (1997). Connectives and narrative text: The role of continuity. *Memory & Cognition*, 25(2):227–236.

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

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.

Riester, A. (2019). Constructing qud trees. In Klaus v. Heusinger, et al., editors, *Questions in Discourse*, volume 2. Brill, Leiden.

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

Rohde, H., Johnson, A., Schneider, N., and Webber, B. (2018). Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sanders, T. J., Demberg, V., Hoek, J., Scholman, M. C., Asr, F. T., Zufferey, S., and Evers-Vermeul, J. (2018).

Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*.

Sanders, T. (2005). Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the exploration and modelling of meaning*, pages 105–114. University of Toulouse-le-Mirail Toulouse.

Segal, E. M., Duchan, J. F., and Scott, P. J. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse processes*, 14(1):27–54.

Siddharthan, A. (2003). Preserving discourse structure when simplifying text. In *Proceedings of the European Natural Language Generation Workshop (ENLG), 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Turnbull, R. (2018). Patterns of probabilistic segment deletion/ reduction in english and japanese. *Linguistics Vanguard*, pages 1–14.

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

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. In *the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1264–1274.

Westera, M. and Rohde, H. (2019). Asking between the lines: elicitation of evoked questions from text. In *Proceedings of the Amsterdam Colloquium*.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31:249–288.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xiong, H., He, Z., Wu, H., and Wang, H. (2019). Modeling coherence for discourse neural machine translation. In *the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 7338–7345.

Yung, F., Scholman, M., and Demberg, V. (2019). Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Linguistic Annotation Workshop LAW*.

9. Language Resource References

Rashmi Prasad and Bonnie Webber and Alan Lee and Aravind Joshi. (2019). *Penn Discourse Treebank version 3.0*. Linguistic Data Consortium, 3.0, ISLRN 977-491-842-427-0.

Ines Rehbein and Scholman Merel and Demberg Vera. (2016). *Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. (Disco-SPICE corpus)*.

Zeyrek, D. and Mendes, A. and Grishina, Y. and Kurfalı, M. and Gibbon, S. and Ogrodniczuk, M. (2018). *TED multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style*.