Annotating QUDs for generating pragmatically rich texts

Christoph Hesse, Anton Benz
Leibniz-Zentrum Allgemeine Sprachwissenschaft
{last name}@zas-berlin.de
Maurice Langner, Felix Theodor, Ralf Klabunde
Ruhr-Universität Bochum, Department of Linguistics
{first name.last name}@rub.de

Abstract

We describe our work on QUD-oriented annotation of driving reports for the generation of corresponding texts – texts that are a mix of technical details of the new vehicle that has been put on the market together with the impressions of the test driver on driving characteristics. Generating these texts pose a challenge since they express non-at-issue and expressive content that cannot be retrieved from a database. Instead these subjective meanings must be justified by comparisons with attributes of other vehicles. We describe our current annotation task for the extraction of the relevant information for generating these driving reports.

1 Introduction

Driving reports about new vehicles, typically published in national daily newspapers and online journals, constitute a text type that poses a challenge for NLG systems since these texts express technical details about these vehicles (often in comparison with previous models or alternative vehicles) combined with subjective impressions of the test driver, resulting in a number of expressive and evaluative expressions. To illustrate these phenomena, we show the English translation of an excerpt from a German driving report about the Porsche Cayenne Turbo S E-Hybrid:

1. *With the Turbo S E-Hybrid strand, Porsche has made a very clever move. The top model, of all models, is no longer the greedy bogeyman, but is ecologically sound when used appropriately. One can of course smile at the statement of the combined consumption of 3.7 litres per 100 km and revile the basis for calculation, but provided that the four-wheel drive car is driven electrically, this value can also be achieved in real terms. Whether this is environmentally friendly or not, especially since electricity is by no means only generated from sun or wind, is another matter.*
(Original text: Mit dem Turbo S E-Hybrid-Strang hat Porsche einen durchaus cleveren Schachzug gemacht. So ist ausgerechnet das Topmodell nicht mehr der gefräßige Buhmann, sondern schlägt sich bei entsprechender Nutzung ökologisch wacker. Man kann die Angabe des kombinierten Verbrauchs von 3,7 Litern je 100 km natürlich belächeln und die Berechnungsgrundlage schmähen, aber unter der Voraussetzung, den Allradler fleißig elektrisch zu fahren, kann dieser Wert auch real zustande kommen. Ob das umweltfreundlich ist oder nicht, zumal Strom bekanntermaßen keineswegs nur aus Sonne oder Wind gewonnen wird, steht auf einem anderen Blatt.)

This text contains facts about consumption and drive type, but most of the text is about subjective estimations and appraisals, realized by evaluative adjectives (*greedy*), adverbs (*of course, by no means*), the use of metaphors (*bogeyman*), and other expressive-related linguistic means. Our research question concerns the relationship between facts and evaluations in driving reports and the justified use of subjective, expressive content in generating these reports.

In generating driving reports, we aim at the explanatory power of *Question under Discussion* (QUD) accounts to text structuring and textual development (van Kuppevelt, 1995; Roberts, 1996). QUD approaches assume that texts are answers to a structured set of explicit or implicit questions. Each QUD does not only impose constraints on the propositional content of a single sentence, but it also determines the focus/background structures and the distinction between at-issue (material that helps to answer the QUD) and non-at-issue content (everything else, typically including evaluative and expressive content). By this, QUD-based approaches provide strong hypotheses about the textual development by successively answering the corresponding QUDs.

These theories provide the starting point of our work: Based on the theory-driven assumptions on textual development, we are going to generate driving reports that are as close to the original texts as possible. If specific content cannot be systematically determined based on QUD requirements, we

get evidence for shortcomings of the underlying theory. Hence, we intend a kind of 'reverse generating' to test the adequacy of QUD-based theories.

The fundamental source for this approach are QUD-annotated data where the QUDs reflect the informational needs to be satisfied by section of the text down to single sentences. It is well known that annotating QUDs in texts requires an intensive training of the annotators and sophisticated annotation guidelines in order to receive reliable results (Arndt Riester and Kuthy, 2018), but if these QUDs have been formulated properly, they provide strict information-structural constraints for their answer that can be used in the generation process. Therefore, we will introduce the underlying data, the problems that occur in annotating the driving reports, and first results concerning resulting QUD structures.

2 Underlying data

We selected 40 driving reports from German online journals (faz.net; welt.de).

The 40 vehicle reports in the corpus were annotated for QUDs and sub-QUDs, focus/background, and non-at-issue content. The guiding principle for us was to be able to separate purely propositional content from non-at-issue, evaluative and expressive content (following Roberts, 1996; Büring, 2003).

QUDs are well-accepted as a discourse-structuring device (Carlson, 1983; van Kuppevelt, 1995; von Stutterheim, 1997). More importantly for us, however, QUDs have been identified as a crucial criterion for distinguishing between at-issue and non-at-issue content. Content which may be non-at-issue with respect to its immediate sub-QUD may nevertheless supply relevant information in the context of an over-arching super-QUD. Analyzing the depth of embedding may give useful insight to enable us to anticipate follow-up QUDs to incomplete sub-QUDs (Onea, 2016).

Crucially for us, QUD approaches recognize that discourse is not merely a coherent presentation of relevant information, but its structure is goal-oriented. Authors consciously decide which questions they want to address, how they want to address them and in what order. Similarly authors make conscious choices about the use of rhetorical relations such as elaboration, contrast, and concession as text-structuring devices, which may or may not be fully predictable from QUD-trees. RST-trees reveal recurring rhetorical structures with predictable evaluative and expressive effect (e.g., employing a contrast relation whenever a vehicle's shortcomings are compensated by other positive attributes, or when comparing a vehicle to its competitors).

We also use QUDs to annotate focus (cf. Roberts, 1996; van Kuppevelt, 1995; von Stutterheim, 1997). Büring (2003) applied Robert's (1996) QUD-stack model to contrastive focus using a QUD-tree model, which under certain conditions can lead to tree structures similar to RST-trees. We use the QUD-trees to arrive at a complete representation of discourse structure. Sub-QUDs are divided further into sub-QUDs until each terminal sub-QUD can be directly mapped to a database entry (e.g., *What is the vehicle's rate of acceleration?* for numerical values or *What type of transmission does it use?* for referring expressions).

The QUD at the root of the tree of each text answers the question whether the vehicle is qualitatively good and worth purchasing. From this root QUD, immediately a sub-QUD derives: Compared to other vehicles or a standard, where the search space is dynamically populated with comparison objects given explicit references in the text (other vehicles in the same class or from competitors) and implicit comparisons based on attribute scales (e.g., vehicles with a comparable type of use or type of engine, transmission, interior, drive assistance features, etc.). We also assume that the typical sections of vehicle reports provide partial answers to the root QUD: a section about the engine, acceleration, gas consumption, mileage, etc. speaks to the quality of the vehicle's performance; a section about the interior speaks to the level of comfort; the test drive speaks to how well the performance promised by the manufacturer holds up under real-world driving conditions, and so forth.

Since an author's subjective view of a vehicle has a tremendous impact on text and information structure, and specifically on the foregrounding and backgrounding of certain information available about the vehicle in the database, capturing authors' subjective evaluation can lead to vastly different QUD phrasings and QUD-structures across annotators. Thus, inter-annotator agreement is the biggest challenge in arriving at systematic discourse structures within this text genre.

The concrete annotation work points to some fundamental problems concerning the assignment

of corresponding tags. Some of them refer to shortcomings of QUD-oriented theories, but others are of a language-specific nature. The main problems that occurred during the annotation process concern so-called feeders (van Kuppevelt, 1995), implicit correlations for contrastive relation, and the assignment of focus.

2.1 Focus

Focus and its complement, background, are information-structural notions that express that part of an utterance that is new to the hearer vs. the information already known by her. In the context of QUD-based theories we could identify that part of a sentence that answers a QUD as focused while the rest belongs to the background.

In many languages focus is expressed phonologically by a lexical item carrying a focus-related accent, the focus exponent. Focus theories explain the position of the focus-related accent by the percolation of a focus feature [F] in a syntactic tree to the bearer of the focus accent. The constituent that expresses the focus, however, is often more extensive that the focus exponent which gives rise to focus ambiguity, i.e. the problem to determine that constituent, given the focus exponent, that expresses the focused information.

For example, the sentence *Anne likes to test the PORSCHE* with *Porsche* being the focus exponent, we have at least two possible focus constituents:

(1) *Anne [likes to test the PORSCHE]$_F$*

(2) *Anne likes to test [the PORSCHE]$_F$*

Which one of these constituents expresses the focus depends on the QUD that shall be answered. Sentence (1) answers the QUD *What's new about Anne?* while the second sentence answers the question *What does Anne like to test?*

In addition to accent placement for expressing focused information, languages provide focus particles and specific syntactic constructions for expressing which part expresses the focus. German (and English) provides all three possibilities, but accent placement is the most prominent means for signaling focus.

The relation between focus and the linguistic means for expressing it seems to be transparent so that annotating focus, once the QUD has been established, should not be a major effort. However, our data indicate some problems in focus assignment that have direct repercussions for the development of the annotation tool and some focus theories as well.

Split-focus: In our data, some sentences have two focus constituents that express one focus together. For example, one QUD in a driving report is *What about the power unit?* The sentence that answers this QUD in the text is:

(3) *In der praktischen Außenhaut des 3,60 kurzen Fünftürers war der Antrieb erstmal kaum zu erkennen.*
'In the practical outer skin of the of the 3.60 short five-door car, the power unit was hardly noticeable at first.'

A plausible assignment of focus is to tag *In der praktischen Außenhaut* ('In the practical outer skin') and *war der Antrieb erstmal kaum zu erkennen* ('the power unit was hardly noticeable at first') as being focused, but not 'the 3.60 short five-door car' since this constituent doesn't provide a part for the answer to the QUD. The consequence of these finding, which have hardly been mentioned in the literature, was to adjust the annotation tool to these phenomena by introducing the possibility to set indexes by the annotators in order to express that both foci belong together.

A further but related phenomenon concerns sentences consisting of two coordinated main clauses, each with its own focus, but answering one QUD:

(4) QUD: *How is the Renault Captur?*
Der Renault Captur [wächst]$_F$ und [verändert seinen Charakter]$_F$.
'The Renault Captur grows and changes its character.'

Is is reasonable to assume two separate foci since this coordination refers to two new aspects of the tested car. Both foci are well motivated by the QUD; they demonstrate that one QUD does not necessarily set up one focus only. Ellipses also indicate that the "one QUD – one focus" default can be violated:

(5) QUD: *How have the aesthetics changed, compared to the old Captur?*
[das sieht scharf trainiert]$_F$ und [angriffslustig aus]$_F$.
'that looks sharply trained and ready to attack.'

The non-elliptic sentence in German would be *das sieht scharf trainiert aus und das sieht angriffslustig aus*, with the prefix *aus* separated from the

prefix verb *aussehen* and remaining in the base position, and the subject plus verb stem inserted in the second clause. The ellipsis forces an index as well for expressing that both foci belong together; otherwise the ellipsis cannot be handled correctly.

The final example illustrates the complexity of focus and background tagging with respect to information-structural considerations. In this example one sentence answers two, actually unrelated, QUDs:

(6) QUD: *How is the interior?*
Der Innenraum macht [einen Sprung in eine neue Zeit]$_F$,
'The interior takes a leap into a new era,'

> QUD: *What will make the new era much more pleasant?*
> *die [mit digitalen Instrumenten, großem Bildschirm, schwebender Mittelkonsole, feineren Oberflächen und adretteren Schaltern]$_F$ deutlich angenehmer wird.*
> 'which becomes much more pleasant with digital instruments, large screen, floating center console, finer surfaces and neater switches'

In (6) one sentence from the driving report answers two QUDs formulated by the annotator. The first one will be answered by the focused constituent in the main clause. In order to motivate the relative clause, a new QUD has been stated and the list of attributes of the car functions as focus.

2.2 Feeder sentences

Another challenging aspect the annotation illustrates is the fact that not every sentence answers a QUD at all. An example for this would be a segment like *Nun war ein größerer Schritt fällig* ('It was time for the next big step'), which shifts the topic of the text in a certain direction but does not provide any information relevant to a QUD. Van Kuppevelt (1995) defines this as "a topicless unit of discourse, [...] or one whose topic is no longer prominent at the moment of questioning". We follow his definition and call these segments (linguistic) feeders. Feeders constitue a trigger for QUDs to arise, but they are not motivated by QUDs themselves. Their status seems to be outside the scope of QUD-based theories.

The example below demonstrates that. Since the given context does not require any information about sale figures of former cars, the segment cannot be motivated by a QUD. However, this new information leads to other QUDs arising, because it provides a set of indeterminacies to which there is no information in the given context:

Feeder: *1,2 Millionen Captur sind seit 2013 verkauft worden.*
'1.2 million Captur have been sold since 2013.'

> QUD: *What about the first generation of Captur?*
> QUD: *What did it look like?*
> *Am Anfang mit trüben Scheinwerfern und viel hartem Kunststoff [...]*
> 'With cloudy headlights and a lot of hard plastic in the beginning [...]'

Feeder sentences often function as an introduction to a new topic, therefore most of them can be found at the start of a new paragraph or unit of text. As van Kuppevelt (1995) notes, even segements that provide information relevant to a QUD can technically act as a feeder as well (if they raise new questions), but we restrict the annotation of feeders to cases in which their appearance is clearly not motivated by a QUD.

2.3 Contrast

QUD approaches emphasize the goal-oriented nature of a text's information structure. Authors' primary goal in the driving report genre is to evaluate a vehicle based on its overall qualities. In order to arrive at an overall evaluation, authors examine inidivual topic areas such as technical specifications, driving experience, comfort, and accessories in turn. Often times authors will note that outstanding performance in one area compensates for deficits in other areas, or that performance in one area is striking compared to previous models or competitors. This makes *Contrast* one of the most common discourse relations found in driving reports, and authors use a variety of surface realizations to express contrast without marking it overtly (no use of the constrastive marker *but*). The following example shows some of these strategies:

1. *Harmonious gliding or hard driving at the limit, the GS, which has become five kilograms heavier, masters both without any efforts. Fortunately, the BMW developers succeeded not only in improving the quality of the exhaust gases, but also in reducing fuel consumption by 0.2 litres/100 km: Despite the fact that the driving style was by no means restrained, the lavishly equipped on-board computer of the test bike showed just 4.8 litres*

per 100 kilometres.
(Original text: Harmonisches Gleiten oder hartes Fahren am Limit, beides beherrscht die um fünf Kilogramm schwerer gewordene GS quasi mit links (π_1). Erfreulicherweise gelang es den BMW-Entwicklern zugleich, nicht nur die Abgasqualitaät zu verbessern, sondern auch den Verbrauch um 0,2 Liter/100 km zu reduzieren (π_2): Trotz keineswegs zurückhaltender Fahrweise zeigte der üppig bestückte Bordcomputer des Testbikes gerade mal 4,8 Liter pro 100 Kilometer an (π_3).)

The evaluative adverb and discourse marker *erstaunlicherweise* (fortunately) marks a *Contrast* relation, but note this relation does not hold between two explicit propositions in the text, rather it holds between (i) the conjoined explicit propositions π_2 and π_3, and (ii) the *unforefilled* implicit expectation of higher gas consumption (and with that poorer exhaust quality), expectations raised by the appositive *um fuenf Kiligramm schwerer gewordene* (weight increase of 5 kg) in π_1. (Simons et al., 2011) claim that appositives are not-at-issue because they do not speak to the QUD answered by the matrix clause which contains the appositive. However, the appositive *um fünf Kilogramm schwerer gewordene* is only *locally* not-at-issue because globally it is very much at-issue for the Contrast relation that follows.

The use of the evaluative adverb and contrastive marker *erstaunlicherweise* (fortunately) is licensed by positive surprisal. Surprisal presupposes a difference between the expected and the actual, and when this difference is positive, i.e. when the actual surpasses the expected, the surprisal is positive and the adverb is licensed. If the new model of the BMW bike consumes 0.2 L/100 km less than the previous model, the previous model consumed 5 L/100 km. Because of the new model's higher weight, its expected gas consumption should be >5 L/100 km. So the surprisal is two-fold: (1) the actual consumption is less compared to the previous model ($4.8 < 5$), and (2) it is less compared to the consumption expected due to the bike being heavier than the previous model ($4.8 < [> 5]$). Since the new model consumes both less than the previous model and less than expected due to weight, *erstaunlicherweise* (fortunately) is double-lincensed. Mentioning the hard driving conditions during testing only emphasizes the level of surprise, while the explicit mention of the onboard computer emphasizes the reliability of the measurements.

Crucially, It is the appositive in π_1 which raises (or at least explicitly adds to) this expectation of higher gas consumption (hard driving conditions $\rightarrow$ higher consumption $\wedge$ higher weight $\rightarrow$ higher consumption). The joint-marker *nicht nur ... sondern auch* (not only ... but also) introduces the two consequents in π_2: gas consumption and exhaust gas quality. The explicitly mentioned weight increase raises causal expectations: higher weight $\rightarrow$ more gas consumption $\rightarrow$ more exhaust gases $\rightarrow$ poorer exhaust gas quality. The contrast relation holds for both the surprisingly good exhaust gas quality and the bike's gas consumption. Both of these implicit contrasts require the assumptions raised by the appositive in π_1. So while the appositive locally may be not-at-issue for how the bike handles, it must be globally at-issue to explain the overtly marked *Contrast* between expected higher gas consumption (and poorer exhaust quality) and the surprisal of actual gas consumption (and exhaust) being lower.

The implicit *Contrast* suggests that the topical QUD of this text should be something like 'Why was the reduction of gasoline consumption surprising/unexpected?' But this would mean the *embedded* appositive needs to be structurally on an equal level with the reduction propositions π_2 and π_3 while the QUD of the matrix proposition in π_1 should be something like 'How does the bike handle under smooth and hard driving conditions?' So while embedding the appositive suggests that the matrix clause's QUD supersedes the appositive's relevant QUD, the *Contrast* relation makes it clear that the QUD hierarchy is actually inverse to the embedding structure. We find this sort of *Contrast* relation with implicit expectations and causal relations raised by technical details quite frequently in our corpus. Our hope is that proper QUD structures which capture the implicit expectations can enrich debates about contrast marking (Jasinskaja and Zeevat, 2008) and information structure (Umbach, 2005).

3 Text planning

Our preliminary analysis of the corpus shows that, broadly speaking, vehicle reports are divided into three parts: (1) an introduction, which may give background information on the manufacturer, occasion for the new release (e.g., anniversaries), stylistic or technical choices characteristic of the vehicle situated in a line of previous models or the history of the line; (2) a main part, which consists of (2a) general technical specifications as advertized by the manufacturer and (2b) impressions from the test drive; (3) an outro which may include price

listings for different models of the vehicle (plus accessories), release dates or additions/changes the manufacturer is planning before the release. Part (2a) usually tends to focus on the most crucial technical details, especially changes which have been made compared to previous models. Part (2b), in stark contrast, is usually a visceral, metaphor- and idiom-rich description of the driving experience aimed at emotionally immersing the reader.

The more engaging these texts are, the more they deviate from this generic structure: Aspects of the vehicles which are exceptionally good or exceptionally bad are foregrounded. We will predict striking features of vehicles via pair-wise feature comparison to other vehicles in the same category. Given a large comparison class, 'average' features will cluster normally around a mean along an evaluation dimension (e.g., less gas consumption is better) while expectable features will correspond to extreme values on either tail of the average distribution. An exceptionally good vehicle excels in all categories that the generic structure explicitly discusses. When a vehicle does not tick all its boxes, authors often restructure the text to make clear how certain excellent features in some categories make up for the shortcomings in other categories. Authors also make a conscious decision to note positive things about a vehicle before diving into its shortcomings, and they try to end on a positive note.

Evaluating the technical details in our database along quality dimensions by comparing vehicles against other vehicles *as well as* comparing different aspects of a vehicle with other aspects of the same vehicle is fundamental for our approach. The overall evaluation made in a vehicle report is the sum of the evaluation of its individual aspects. Not all technical details contained in the ADAC database are explicitly mentioned in the reports, and of those that are mentioned, some are given more weight than others in contributing to the overall evaluation. We aim to derive this weighting probabilistically from the comparison analysis of technical features in the database. Since quality dimensions associated with these features are subjective, these are based on the original annotation.

The type of vehicle (e.g., ICE, internal combustion engine, versus EV, electric vehicle; car versus motorbike) pre-selects a subset of relevant technical features as well as a class-specific document plan. We then go through the evaluation process as described above. The result of this process is a linearized text plan with vehicle features weighted for relevance and impact on the overall evalutation. We assume that non-at-issue content does not directly answer a proposition's immediate QUD, but, instead, it contextualizes the choices authors make in establishing the foregrounding and backgrounding of vehicle features and marks subjective evaluation. With the evaluation process complete, the text plan can be enriched with non-at-issue content.

4 Surface realisation

For realizing the sentences, a hierarchy of classes has been set up which defines messages for categorical pieces of information that are stereotypically produced in the genre of vehicle reports, e.g. 'HorsePowerMSG'. Each of these classes may perform its own lexicalisation task by a proper interface function. A microplanner class provides containers for messages, on which aggregation tasks and other post-processing may operate.

Among those post-processors, a module for referring expression generation and coreference realization are going to be implemented. Across the microplan, references to the object under discussion are filled with placeholders. A suitable method for this is based on the QUD structure and the depth of embedding of paragraphs, which limit the availability of entities and prevents the usage of pronominal reference. A focus-stack model keeps track of mentioned entities and the different lexicalisation options for the object at the given position.

A lexicon is built from the corpus including idioms, which allows for a probabilistic distribution over head verbs that may be used to lexicalize different messages. The subcategorization frames allow to further process both syntactic and morphologic processes.

Simplifications must be made according to background knowledge and authors' opinions regarding different cars, which would demand for a complex common sense reasoning database. Instead, we intend to use predefined templates for these portions of text in order to achieve our aim of showing whether this non-at-issue content can be generated.

For surface realisation, we use the Java library of SimpleNLG for German (Bollmann, 2011), which covers mainly morphological operations. An interface between micro-planning and SimpleNLG is needed in order to call the correct methods for the respective syntactic constructions defined by the

micro-planner. This means that an interpreter for the micro-planning implements both a linearization of lexemes and a mapping from AVM structure to Java methods in SimpleNLG.

5 Summary and outlook

Without annotations with sufficient quality, one cannot generate good texts. We are interested in adopting the QUD-approach to text structuring to generating reports in order to test the soundness of this approach. QUD-based linguistic analyses tend to be confined to simplified texts with a focus on relevant phenomena; we want to know whether such a theory-driven approach to generating pragmatically rich texts is feasible.

References

Liesa Brunetti Arndt Riester and Kodula De Kuthy. 2018. *Annotation guidelines for Questions under Discussion and information structure*, pages 403–443. Benjamins, Amsterdam.

Marcel Bollmann. 2011. Adapting SimpleNLG to German. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011)*, page 133–138, Nancy.

Daniel Büring. 2003. On D-trees, beans, and B-accents. *Linguistics & Philosophy*, 26(5):511–545.

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

Katja Jasinskaja and Hank Zeevat. 2008. Explaining additive, adversative and contrast marking in russian and english. *Revue de Sémantique et Pragmatique*, 24:65–91.

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

Edgar Onea. 2016. *Potential Questions at the Semantics-Pragmatics Interface*, volume 33 of *Current Research in the Semantics/Pragmatics Interface*. Brill, Leiden.

Craige Roberts. 1996. Information structure in discourse: Toward an integrated formal theory of pragmatics. In Jar Hak Yoon and Andreas Kathol, editors, *OSU Working Papers in Linguistics*, volume 49, pages 91–136. The Ohio State University, Department of Linguistics, Ohio.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2011. What projects and why? *Semantics and Linguistic Theory*, 20:309–327.

Christiane von Stutterheim. 1997. *Einige Prinzipien des Textaufbaus: Empirische Untersuchungen zur Produktion mündlicher Texte*, volume 184 of *Reihe Germanistische Linguistik*. Niemeyer Verlag, Tübingen.

Carla Umbach. 2005. Contrast and information structure: A focus-based analysis of *but*. *Journal of Linguistics*, 43:207–232.