Variability in Annotations of Questions under Discussion

Geraldine Baumann (Ruhr University Bochum), Tatjana Scheffler (Ruhr University Bochum)
ESSLLI 2026 (Prague), Workshop on Human Label Variation in Discourse and Pragmatic Phenomena, 10.08.26

RUB

SFB 1287 Limits of Variability in Language
Cognitive, Computational, and Grammatical Aspects

What are QUDs?

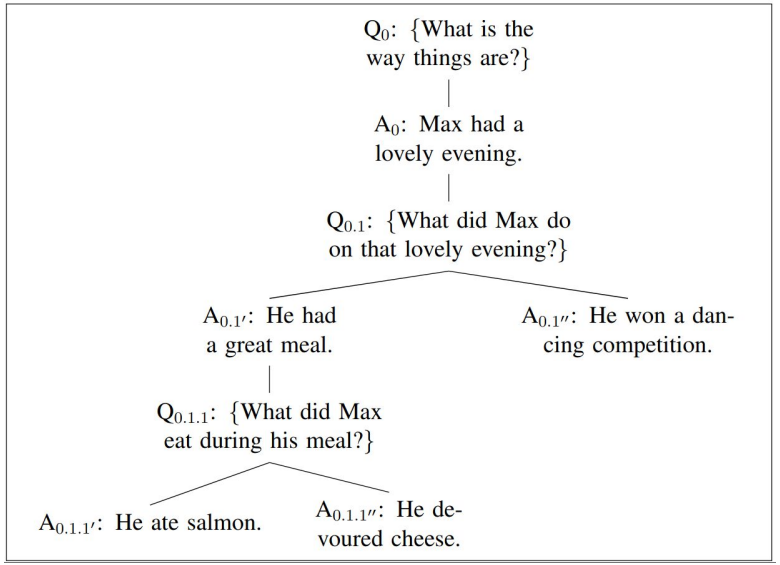
What are *Questions under Discussion*?

- introduced by Roberts (1996/2012) analyses discourse through ordered (explicit or implicit) current questions raised by the context

- Aim: Answering the *"big question"*: What is the way things are? Subquestions create hierarchical (tree) structure.

- very nuanced annotations with visible subtle differences due to 'free form' annotations; however, also highly subjective and susceptible to HLV

π1: Max had a lovely evening.
π2: He had a great meal.
π3: He ate salmon.
π4: He devoured cheese.
π5: He won a dancing competition.



Riester (2019)

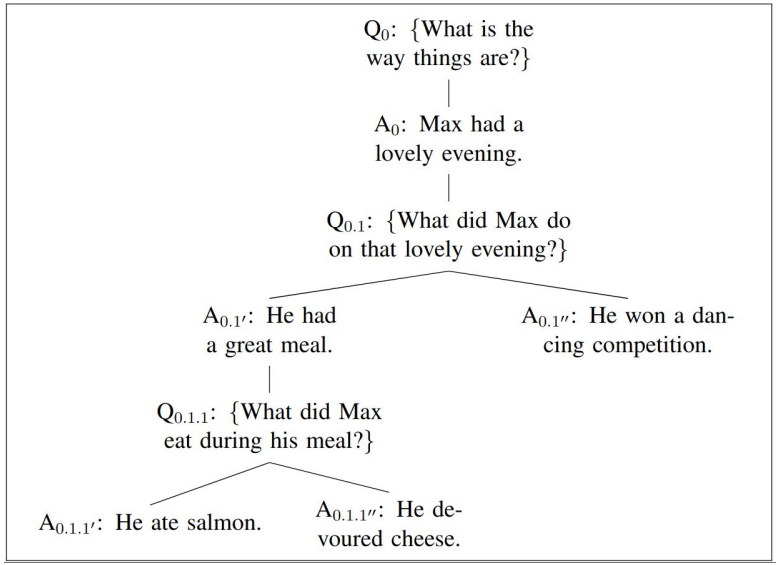
How are QUDs annotated?

Most well known and extensive guidelines are by Riester et al. (2018):

- for each text segment a QUD is formulated following three principles:

- Q-A-Congruence:
  “QUDs must be answerable by the assertion(s) that they immediately dominate.”
- Maximize-Q-Anaphoricity:
  “Implicit QUDs should contain as much given material as possible.”
- Q-Givenness:
  “Implicit QUDs can only consist of given (or, at least, highly salient) material.”

π1: Max had a lovely evening.
π2: He had a great meal.
π3: He ate salmon.
π4: He devoured cheese.
π5: He won a dancing competition.



Riester (2019)

How to compare QUD annotations?

- QUDs mostly compared in terms of structure and segmentation (e.g. De Kuthy et al. 2018, Namuduri et al. (2025), Wu et al. (2023)); approaches largely disregard QUD similarity (Deck et al. upcoming)
- currently no good method for comparing QUD annotations (assessing parses)

Deck et al. (upcoming) developed a method for integrating question similarity analysis into QUD annotation comparisons → only tested on two text types and with a gold standard segmentation

How to compare QUD annotations?

- QUDs mostly compared in terms of structure and segmentation (e.g. De Kuthy et al. 2018, Namuduri et al. (2025), Wu et al. (2023)); approaches largely disregard QUD similarity (Deck et al. upcoming)
- currently no good method for comparing QUD annotations (assessing parses)

Deck et al. (upcoming) developed a method for integrating question similarity analysis into QUD annotation comparisons → only tested on two text types and with a gold standard segmentation

Aim: Extend the method proposed by Deck et al. (upcoming) for annotation with different segmentations

Comparing QUD Annotations

Methodology

Comparing QUDs (Deck et al. upcoming):

	Similar QUDs
Prior	She picks the lock to Mrs Simpson’s room.
QUD	• Has Susie been to Mrs Simpson's room before? • Has Susie been in the room before? • Has Susie ever been in Mrs Simpson's room before?
Post	Susie’s in there every week to clean
Sim. all-mpnet-base-v2	0.89

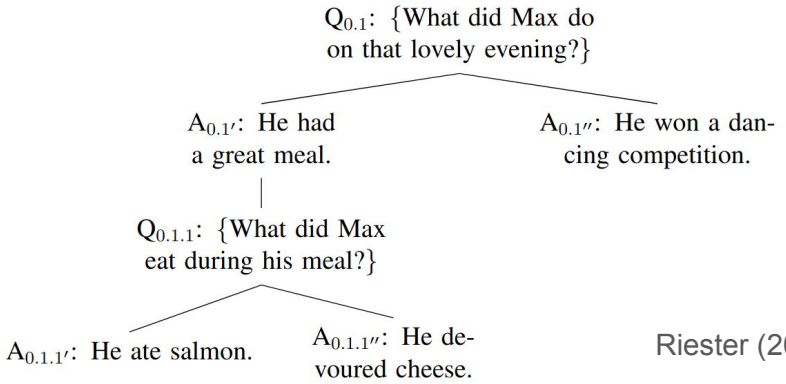
Comparing QUDs (Deck et al. upcoming):

	Similar QUDs	Dissimilar QUDs
Prior	She picks the lock to Mrs Simpson's room.	A single flake of dead skin can ruin a microchip.
QUD	• Has Susie been to Mrs Simpson's room before? • Has Susie been in the room before? • Has Susie ever been in Mrs Simpson's room before?	• What about the factory? • Why is scrubbing her face an issue for Susie? • How does that constitute a discomfit?
Post	Susie's in there every week to clean	But the factory is also vacuum-dry and she (Susie) has to moisturise often.
Sim. all-mpnet-base-v2	0.89	0.08

How to compare QUD annotations?

Aim: adjust method developed by Deck et al. (upcoming) for annotations with differing segmentations and use it on a larger dataset

- since QUD trees follow a hierarchical structure, all segments in a given branch are (partial) answers to the governing question → allows us to align segmentations



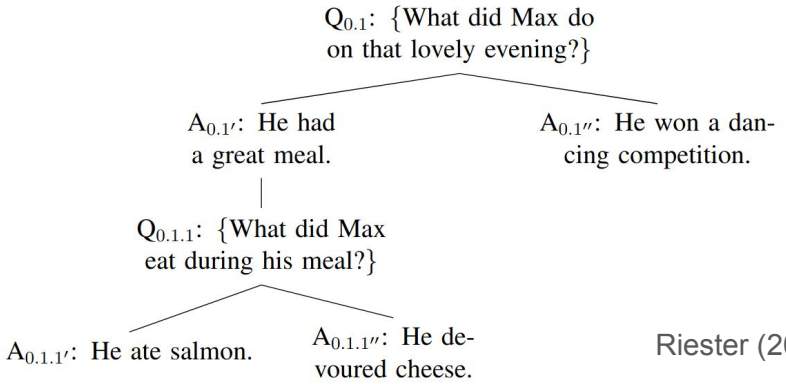
Riester (2019)

How to compare QUD annotations?

Aim: adjust method developed by Deck et al. (upcoming) for annotations with differing segmentations and use it on a larger dataset

- since QUD trees follow a hierarchical structure, all segments in a given branch are (partial) answers to the governing question → allows us to align segmentations

- create unified (gold standard) segmentation
- assign each segment to the closest preceding QUD within the same discourse branch
- compare QUDs across annotations using sentence embeddings (Deck et al. (upcoming))



Riester (2019)

Data

QUD-Anno-Challenge:

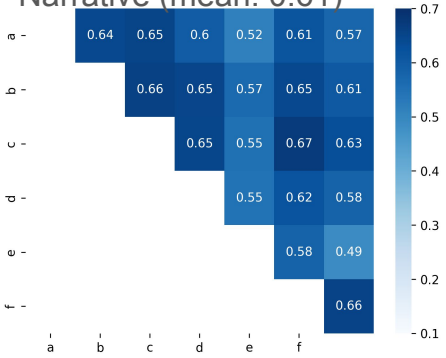
- three long-form texts from three different genres (narrative, interview, review) were annotated by multiple teams
- no gold segmentation between teams, partially within team

	Narrative (written)	Interview (spoken)	Review (written)
Sentences	77	42	58
Tokens	964	804	1131
Types	437	309	511
Type-Token Ratio (TTR)	0.453	0.384	0.452
Avg. Sentence Length	12.52	19.14	19.50
Annotations	7	7	2
Same Segmentations	4	4	0
Segments (aligned)	98	48	76

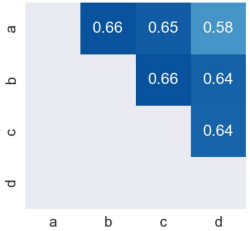
Similarities Results

Question Similarity Overview:

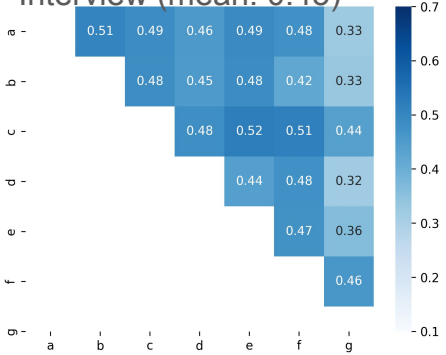
Narrative (mean: 0.61)



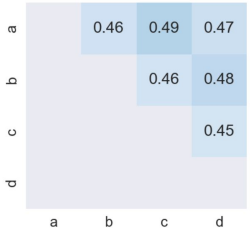
Deck et al. (upcoming): mean 0.64



Interview (mean: 0.45)



Deck et al. (upcoming): mean 0.47



Pairwise Annotator Similarities (all-mpnet-base-v2):

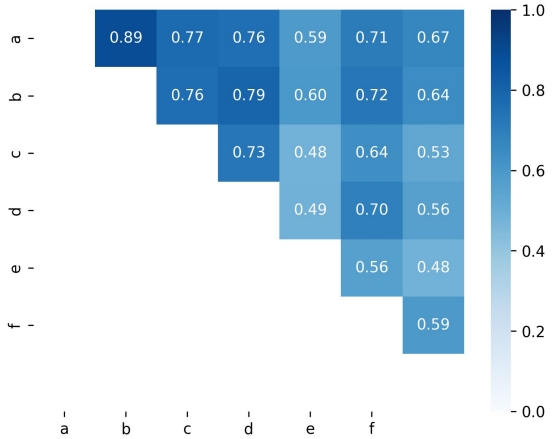
- Review Mean: 0.43 (only two annotators with different segmentations)
- for the narrative text and the interview, annotators *a* to *d* had the same segmentations
- also shows higher similarity scores for narrative text than for interview and review
- annotator *g* is a QUD expert (lower sim. in interview)

Tree Structure Similarity Overview:

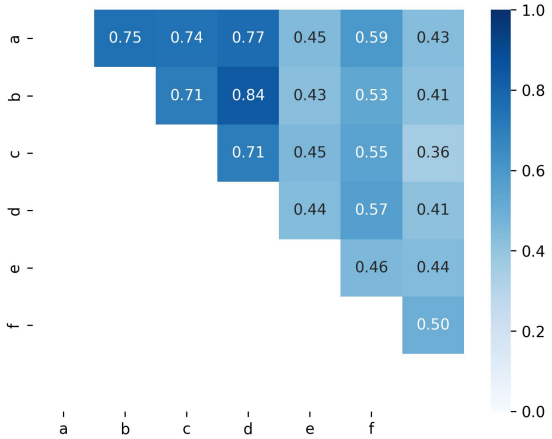
Also compared structure of the trees: token-based pairwise Parseval Score (F1) (Manning and Schütze (1999)

- Review mean: 0.52
- narrative again shows the highest results
- Deck et al. (upcoming): Narrative mean = 0.79, Interview mean = 0.73

Narrative (mean: 0.65):



Interview (mean: 0.55):



	Narrative	Interview	Review
Boundary Agreement (Fournier 2013)	0.73	0.64	0.68
Crossing Accuracy:	0.87	0.81	0.76

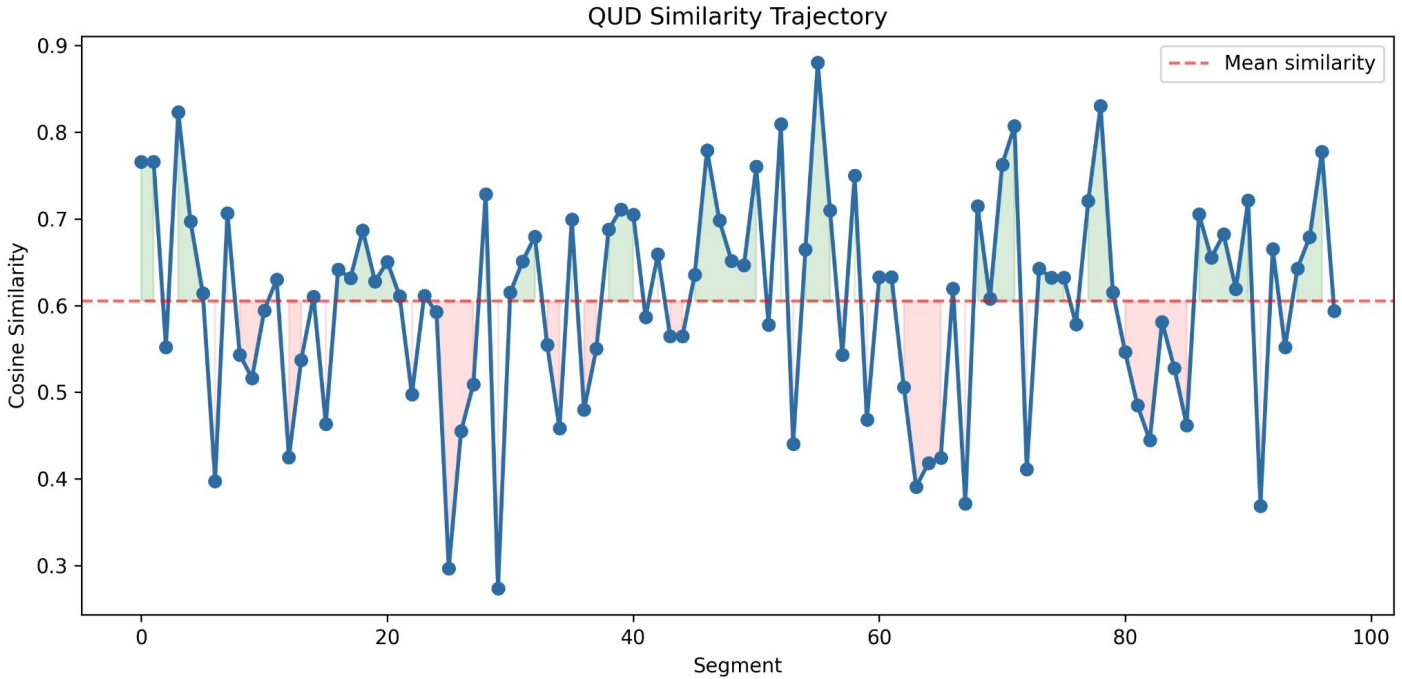
Similarity Trajectory

Where do QUD annotations differ?

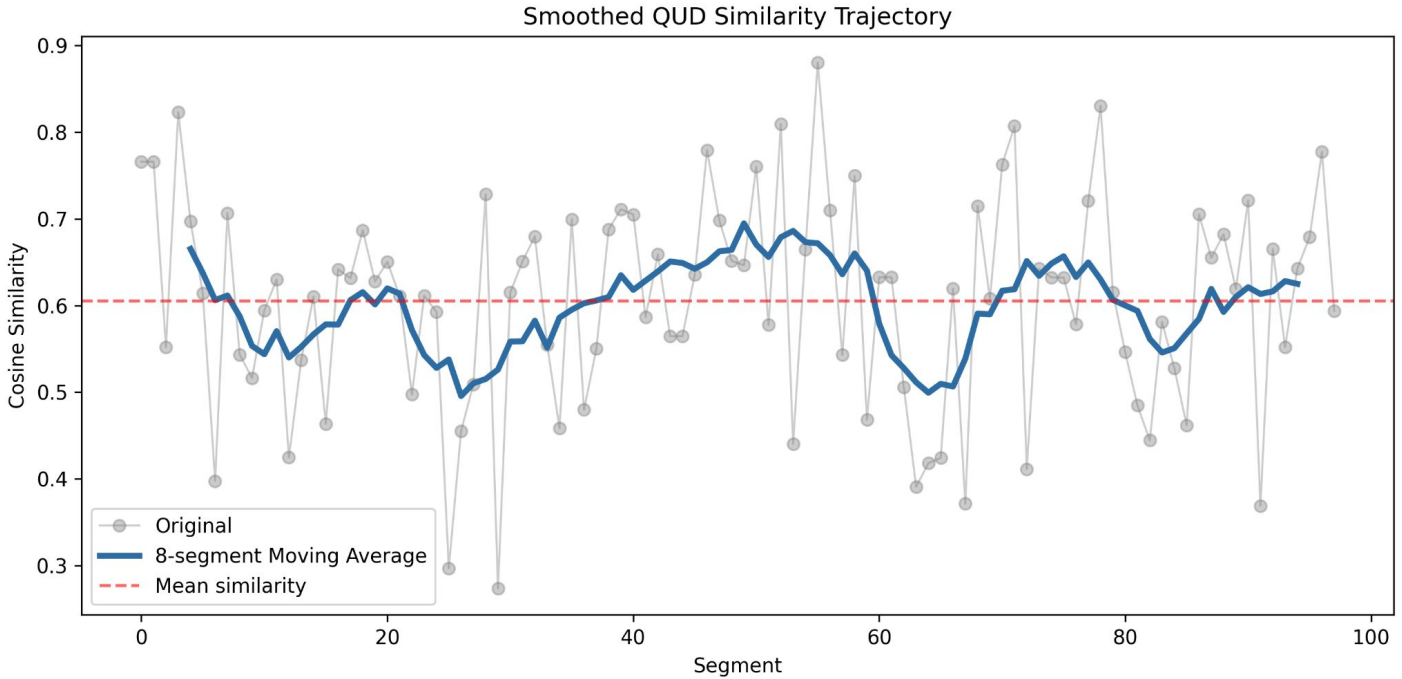
- trying to identify regions of highly different/similar QUDs within the text: plotting ‘trajectory’ over text length by → see where the similarity ‘dips’ or ‘spikes’
- areas can later be examined for potential reasons for changes in similarity

- visualisations as a line diagram
- smoothed using moving average (window = 8)
- if more than two annotators: mean similarity of all pairwise similarities used

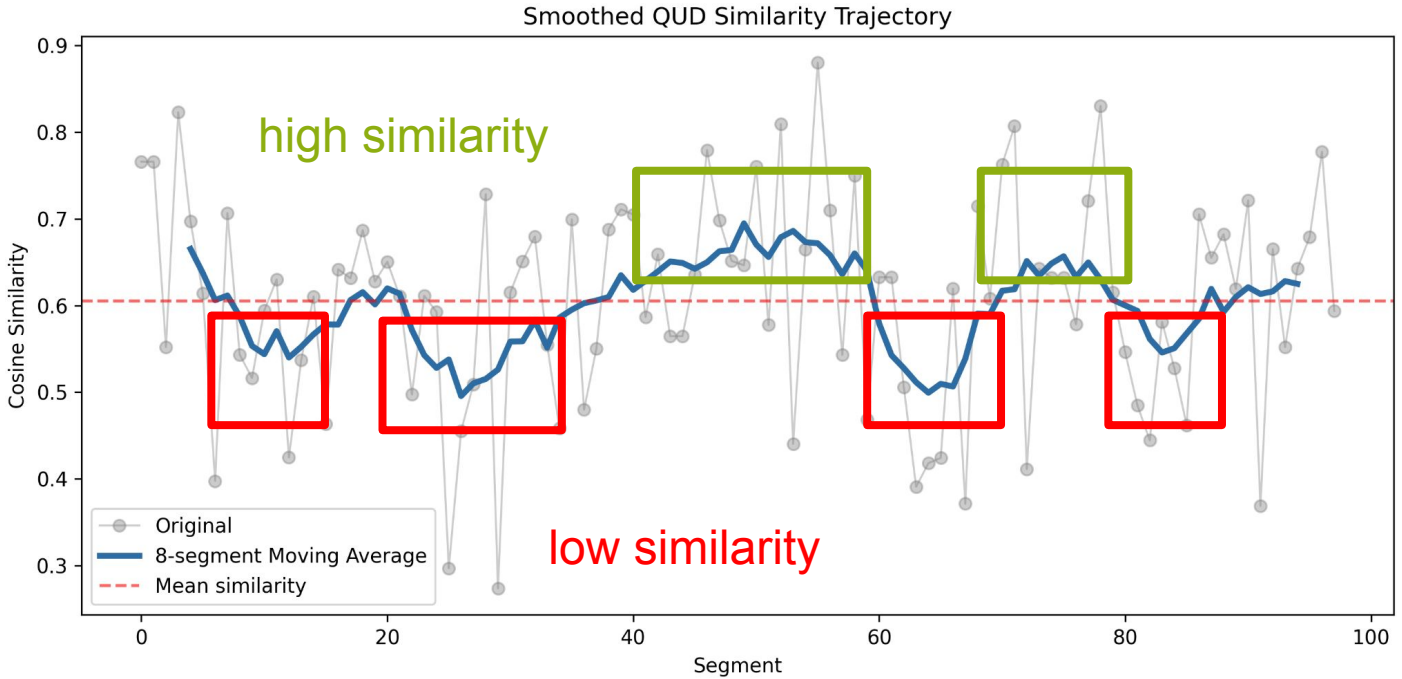
QUD Similarity Trajectory (Narrative Text Example)



QUD Similarity Trajectory (Narrative Text Example)

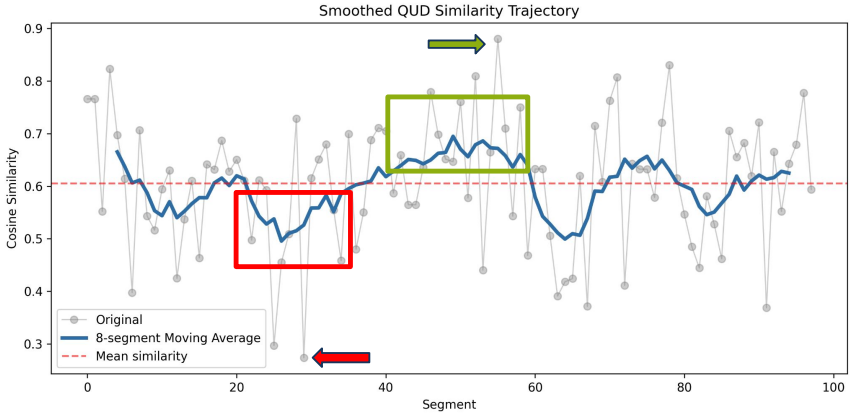


QUD Similarity Trajectory (Narrative Text Example)



Example

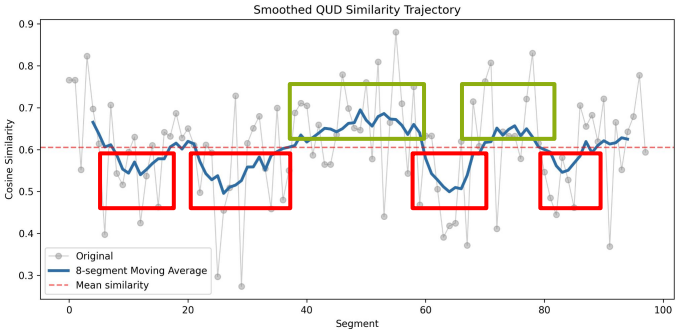
- in both example cases (high and low similarity) many of the closely preceding segments have comparable similarity scores
- could hint at different / very similar interpretations of the discourse in these areas (instead of a single ambiguous segment)



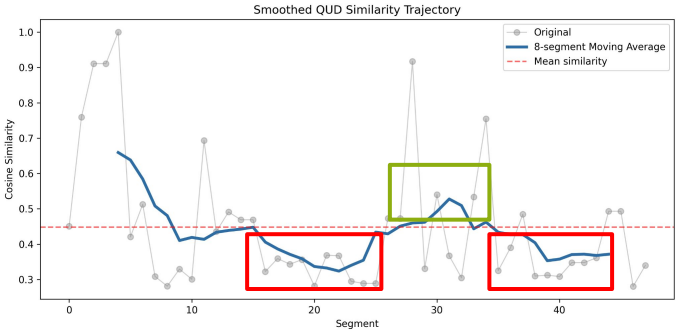
Prior	She picks the lock to Mrs Simpson's room.	A single flake of dead skin can ruin a microchip.
QUD	• Has Susie been to Mrs Simpson's room before? • Has Susie been in the room before? • Has Susie ever been in Mrs Simpson's room before?	• What about the factory? • Why is scrubbing her face an issue for Susie? • How does that constitute a discomfit?
Sim.	0.89	0.08

QUD Similarity Trajectory Comparison

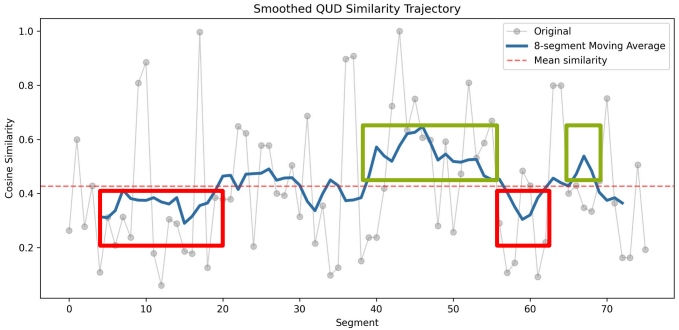
Narrative:



Interview:



Review:



- especially interview many low ranking QUDs (despite more explicit questions): difficult spoken structure, evasive responses (genre-specific)

Takeaways

Takeaways:

- alignment: thanks to the structure of QUD annotations, simple hierarchy-based alignment to separate gold standard segmentations already provides good results,
  but still looses information → test more alignment methods
- importance of genre: texts following a narrative structure seem to provide very similar QUD annotations, while interviews, despite the amount of explicit questions, show low similarities → test on more texts and different genres
- structural difficulties: we can identify areas of low and high similarity by smoothing similarity trajectories, which can help identify areas where the discourse structure is particularly difficult → identify reasons for variability

Thank you!

Works Cited

- Deck, Oliver, Tatjana Scheffler, and Hannah J. Seemann (upcoming). QUD Structure as Discourse Structure: Quantitative and Qualitative Comparison of Annotations. In preparation.
- De Kuthy, Kordula, Nils Reiter, and Arndt Riester (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). https://aclanthology.org/L18-1304/
- Fournier, Chris (2013). Evaluating text segmentation using boundary edit distance. Hinrich Schuetze, Pascale Fung & Massimo Poesio (eds.), *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers),* 1702–1712. https://aclanthology.org/P13-1167
- Manning, Christopher D. and Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. MIT Press.
- Namuduri, Ramya, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li (2025). QUDSIM: Quantifying Discourse Similarities in LLM-Generated Text. DOI: doi.org/10.48550/arXiv.2504.09373
- Riester, A. (2019). Constructing QUD Trees. In: M. Zimmermann, K. von Heusinger, & V. E. Onea Gaspar (Eds.), *Questions in Discourse* (pp. 164–193). Brill. https://doi.org/10.1163/9789004378322_007
- Riester, A., Brunetti, L., & Kuthy, K. (2018). Annotation Guidelines for Questions under Discussion and Information Structure. *Studies in Prosody and Syntax*. 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. https://doi.org/10.3765/sp.5.6
- Shahmohammadi, Sara, Hannah Seemann, Manfred Stede, and Tatjana Scheffler (2023). Encoding discourse structure: Comparison of RST and QUD. *Proceedings of the 4th Workshop on Computational Approaches to Discourse* (CODI 2023). https://doi.org/10.18653/v1/2023.codi-1.11
- Wu, Yating, Ritika Mangla, Greg Durrett, and Junyi Jessy Li (2023). QUDeval: The Evaluation of Questions Under Discussion Discourse Parsing. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. DOI: doi.org/10.18653/v1/2023.emnlp-main.325