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

Lisa Schäfer, Robin Lemke, Bozhidara Hristova, Heiner Drenhaus, Ingo Reich

Project B3, SFB 1102, Saarland University, Saarbrücken

The QUD-Anno Challenge

23 February 2023

IDeaL SFB 1102

UNIVERSITÄT DES SAARLANDES

Motivation

Susie lifts the lid of the abandoned teapot

Motivation

Susie lifts the lid of the abandoned teapot

Motivation

Susie lifts the lid of the abandoned teapot

What is happening?

Motivation

Susie lifts the lid of the abandoned teapot

What is happening?
What does Susie do?

Motivation

Susie lifts the lid of the abandoned teapot

What is happening?
What does Susie do?
What does Susie lift the lid of?
…

Motivation

(1) Susie lifts the lid of the abandoned teapot

Possible QUDs = {*What is happening?*
What does Susie do?
What does Susie lift the lid of?
...}

Motivation

(1) Susie lifts the lid of the abandoned teapot

Possible QUDs = {*What is happening?*
What does Susie do?
What does Susie lift the lid of?
...}

Which Question Under Discussion does (1) answer?

Motivation

(1) Susie lifts the lid of the abandoned teapot

Possible QUDs = {*What is happening?*
What does Susie do?
What does Susie lift the lid of?
...}

Which Question Under Discussion does (1) answer?

⇒ Probability distribution over possible QUDs based on crowdsourced questions

Motivation

(1) Susie lifts the lid of the abandoned teapot

Possible QUDs = {*What is happening?* 48%
What does Susie do? 21%
What does Susie lift the lid of? 8%
…}

Which Question Under Discussion does (1) answer?

⇒ Probability distribution over possible QUDs based on crowdsourced questions

Overview

Overview

Background

- Assumption: each assertion in a text answers one Question Under Discussion

(von Stutterheim and Klein, 1989; van Kuppevelt, 1995; Roberts, 2012)

$\Rightarrow$ Previous research: QUDs annotated by experts using elaborate guidelines

(De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019)

Overview

Background

- Assumption: each assertion in a text answers one Question Under Discussion

(von Stutterheim and Klein, 1989; van Kuppevelt, 1995; Roberts, 2012)

⇒ Previous research: QUDs annotated by experts using elaborate guidelines

(De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019)

Our contribution (cf. Westera et al., 2020; Poppels and Kehler, to appear; Reich et al., to appear)

Overview

Background

- Assumption: each assertion in a text answers one Question Under Discussion

(von Stutterheim and Klein, 1989; van Kuppevelt, 1995; Roberts, 2012)

⇒ Previous research: QUDs annotated by experts using elaborate guidelines

(De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019)

Our contribution (cf. Westera et al., 2020; Poppels and Kehler, to appear; Reich et al., to appear)

- Data set of crowdsourced non-expert annotations of QUDs to investigate:

Overview

Background

- Assumption: each assertion in a text answers one Question Under Discussion

(von Stutterheim and Klein, 1989; van Kuppevelt, 1995; Roberts, 2012)

⇒ Previous research: QUDs annotated by experts using elaborate guidelines

(De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019)

Our contribution (cf. Westera et al., 2020; Poppels and Kehler, to appear; Reich et al., to appear)

- Data set of crowdsourced non-expert annotations of QUDs to investigate:

❶ Which QUDs do naïve comprehenders actually assume when processing texts?

Overview

Background

- Assumption: each assertion in a text answers one Question Under Discussion

(von Stutterheim and Klein, 1989; van Kuppevelt, 1995; Roberts, 2012)

⇒ Previous research: QUDs annotated by experts using elaborate guidelines

(De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019)

Our contribution (cf. Westera et al., 2020; Poppels and Kehler, to appear; Reich et al., to appear)

- Data set of crowdsourced non-expert annotations of QUDs to investigate:

❶ Which QUDs do naïve comprehenders actually assume when processing texts?

❷ (To what extent) does the distribution over possible QUDs vary

Overview

Background

- Assumption: each assertion in a text answers one Question Under Discussion

(von Stutterheim and Klein, 1989; van Kuppevelt, 1995; Roberts, 2012)

⇒ Previous research: QUDs annotated by experts using elaborate guidelines

(De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019)

Our contribution (cf. Westera et al., 2020; Poppels and Kehler, to appear; Reich et al., to appear)

- Data set of crowdsourced non-expert annotations of QUDs to investigate:

❶ Which QUDs do naïve comprehenders actually assume when processing texts?

❷ (To what extent) does the distribution over possible QUDs vary
  - across the course of a text?

Overview

Background

- Assumption: each assertion in a text answers one Question Under Discussion

(von Stutterheim and Klein, 1989; van Kuppevelt, 1995; Roberts, 2012)

⇒ Previous research: QUDs annotated by experts using elaborate guidelines

(De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019)

Our contribution (cf. Westera et al., 2020; Poppels and Kehler, to appear; Reich et al., to appear)

- Data set of crowdsourced non-expert annotations of QUDs to investigate:

❶ Which QUDs do naïve comprehenders actually assume when processing texts?

❷ (To what extent) does the distribution over possible QUDs vary
  - across the course of a text?
  - between texts of different genres?

Overview

Background

- Assumption: each assertion in a text answers one Question Under Discussion

(von Stutterheim and Klein, 1989; van Kuppevelt, 1995; Roberts, 2012)

⇒ Previous research: QUDs annotated by experts using elaborate guidelines

(De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019)

Our contribution (cf. Westera et al., 2020; Poppels and Kehler, to appear; Reich et al., to appear)

- Data set of crowdsourced non-expert annotations of QUDs to investigate:

❶ Which QUDs do naïve comprehenders actually assume when processing texts?

❷ (To what extent) does the distribution over possible QUDs vary
  - across the course of a text?
  - between texts of different genres?

❸ (To what extent) can non-expert annotations complement expert annotations?

Overview and outline

Overview and outline

1. **Segmentation** of source texts into atomic assertions

Overview and outline

❶ **Segmentation** of source texts into atomic assertions

❷ **Production task**: Collection of ~ 30 questions per assertion (crowdsourced)

Overview and outline

❶ **Segmentation** of source texts into atomic assertions

❷ **Production task**: Collection of ~ 30 questions per assertion (crowdsourced)

❸ **Preprocessing** and **filtering** of produced questions

Overview and outline

❶ **Segmentation** of source texts into atomic assertions

❷ **Production task**: Collection of ~ 30 questions per assertion (crowdsourced)

❸ **Preprocessing** and **filtering** of produced questions

❹ **Annotation**: Pooling of semantically identical questions by expert annotators

Overview and outline

❶ **Segmentation** of source texts into atomic assertions

❷ **Production task**: Collection of ~ 30 questions per assertion (crowdsourced)

❸ **Preprocessing** and **filtering** of produced questions

❹ **Annotation**: Pooling of semantically identical questions by expert annotators

❺ **Annotated data set** with likelihood of QUDs and entropy per assertion

Overview and outline

1. **Segmentation** of source texts into atomic assertions
2. **Production task**: Collection of ~ 30 questions per assertion (crowdsourced)
3. **Preprocessing** and **filtering** of produced questions
4. **Annotation**: Pooling of semantically identical questions by expert annotators
5. **Annotated data set** with likelihood of QUDs and entropy per assertion
6. **Data set statistics**: Distribution of probability and entropy

Data collection

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review

Susie lifts the lid of the abandoned teapot and swirls the water. The teabag sloshes against the sides. The tea is cold and bitter, but Susie doesn't mind because her landlady, Mrs Simpson, normally reuses tea bags. Usually, by the time Susie gets home, the tea mostly tastes of chlorine.

As she checks Mrs Simpson's calendar, Susie rubs the place where the elastic cap from work scrunched all day. A play. The skin feels puckered and soft, like white and wrinkly fingertips in the bath. Her mother used to read beside the clawfoot bathtub her father imported from England. Susie—who wasn't a Susie at all then—would tuck her chin over the edge of the tub and listen. The water would go cold. Her skin would loose and crinkle.

Mrs Simpson only makes fresh tea for Mr Johnson next door. One cup still contains a moss-smoke slick of whiskey. Susie wipes the rim of the cup with her sleeve then pours herself some tea.

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into **atomic assertions** (De Kuthy et al., 2018):

Susie lifts the lid of the abandoned teapot and swirls the water. The teabag sloshes against the sides. The tea is cold and bitter, but Susie doesn't mind because her landlady, Mrs Simpson, normally reuses tea bags. Usually, by the time Susie gets home, the tea mostly tastes of chlorine.

As she checks Mrs Simpson's calendar, Susie rubs the place where the elastic cap from work scrunched all day. A play. The skin feels puckered and soft, like white and wrinkly fingertips in the bath. Her mother used to read beside the clawfoot bathtub her father imported from England. Susie—who wasn't a Susie at all then—would tuck her chin over the edge of the tub and listen. The water would go cold. Her skin would loose and crinkle.

Mrs Simpson only makes fresh tea for Mr Johnson next door. One cup still contains a moss-smoke slick of whiskey. Susie wipes the rim of the cup with her sleeve then pours herself some tea.

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into **atomic assertions** (De Kuthy et al., 2018):
  - each declarative utterance delimited by periods, colons and semicolons

```
Susie lifts the lid of the abandoned teapot and swirls the water.
The teabag sloshes against the sides.
The tea is cold and bitter, but Susie doesn't mind because her landlady, Mrs Simpson,
normally reuses tea bags.
Usually, by the time Susie gets home, the tea mostly tastes of chlorine.
As she checks Mrs Simpson’s calendar, Susie rubs the place where the elastic cap from work
scrunched all day.
A play.
The skin feels puckered and soft, like white and wrinkly fingertips in the bath.
Her mother used to read beside the clawfoot bathtub her father imported from England.
Susie—who wasn't a Susie at all then—would tuck her chin over the edge of the tub and
listen.
The water would go cold.
Her skin would loose and crinkle.
Mrs Simpson only makes fresh tea for Mr Johnson next door.
One cup still contains a moss-smoke slick of whiskey.
Susie wipes the rim of the cup with her sleeve then pours herself some tea.
```

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into **atomic assertions** (De Kuthy et al., 2018):
  - each declarative utterance delimited by periods, colons and semicolons (including fragments (Morgan, 1973))

```
Susie lifts the lid of the abandoned teapot and swirls the water.
The teabag sloshes against the sides.
The tea is cold and bitter, but Susie doesn't mind because her landlady, Mrs Simpson,
normally reuses tea bags.
Usually, by the time Susie gets home, the tea mostly tastes of chlorine.
As she checks Mrs Simpson’s calendar, Susie rubs the place where the elastic cap from work
scrunched all day.
A play.
The skin feels puckered and soft, like white and wrinkly fingertips in the bath.
Her mother used to read beside the clawfoot bathtub her father imported from England.
Susie—who wasn't a Susie at all then—would tuck her chin over the edge of the tub and
listen.
The water would go cold.
Her skin would loose and crinkle.
Mrs Simpson only makes fresh tea for Mr Johnson next door.
One cup still contains a moss-smoke slick of whiskey.
Susie wipes the rim of the cup with her sleeve then pours herself some tea.
```

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into **atomic assertions** (De Kuthy et al., 2018):
  - each declarative utterance delimited by periods, colons and semicolons (including fragments (Morgan, 1973))
  - conjuncts of clausal and verbal coordinations (answer independent QUDs)

```
Susie lifts the lid of the abandoned teapot
and swirls the water.
The teabag sloshes against the sides.
The tea is cold and bitter,
but Susie doesn't mind
because her landlady, Mrs Simpson, normally reuses tea bags.
Usually, by the time Susie gets home, the tea mostly tastes of chlorine.
As she checks Mrs Simpson’s calendar, Susie rubs the place where the elastic cap from work
scrunched all day.
A play.
The skin feels puckered and soft, like white and wrinkly fingertips in the bath.
Her mother used to read beside the clawfoot bathtub her father imported from England.
Susie—who wasn't a Susie at all then—would tuck her chin over the edge of the tub
and listen.
The water would go cold.
Her skin would loose
and crinkle.
Mrs Simpson only makes fresh tea for Mr Johnson next door.
One cup still contains a moss-smoke slick of whiskey.
Susie wipes the rim of the cup with her sleeve
then pours herself some tea.
```

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into **atomic assertions** (De Kuthy et al., 2018):
  - each declarative utterance delimited by periods, colons and semicolons (including fragments (Morgan, 1973))
  - conjuncts of clausal and verbal coordinations (answer independent QUDs)

```
[Susie lifts the lid of the abandoned teapot]A
and [swirls the water]A
The teabag sloshes against the sides.
The tea is cold and bitter,
but Susie doesn't mind
because her landlady, Mrs Simpson, normally reuses tea bags.
Usually, by the time Susie gets home, the tea mostly tastes of chlorine.
As she checks Mrs Simpson’s calendar, Susie rubs the place where the elastic cap from work
scrunched all day.
A play.
The skin feels puckered and soft, like white and wrinkly fingertips in the bath.
Her mother used to read beside the clawfoot bathtub her father imported from England.
Susie—who wasn't a Susie at all then—would tuck her chin over the edge of the tub
and listen.
The water would go cold.
Her skin would loose
and crinkle.
Mrs Simpson only makes fresh tea for Mr Johnson next door.
One cup still contains a moss-smoke slick of whiskey.
Susie wipes the rim of the cup with her sleeve
then pours herself some tea.
```

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into atomic assertions
- Restriction to first 20 assertions of narrative and first 19 assertions of car review

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into atomic assertions
- Restriction to first 20 assertions of narrative and first 19 assertions of car review
- Presentation of title, author and non-declarative utterances (i.e. interrogatives and imperatives from car review) for completeness

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into atomic assertions
- Restriction to first 20 assertions of narrative and first 19 assertions of car review
- Presentation of title, author and non-declarative utterances (i.e. interrogatives and imperatives from car review) for completeness

You will read the beginning of **What Mrs Simpson Knows About Immigrants** by Kinneson Lalor.

Data collection through online production experiment

Materials

- Focus on two of the three provided texts: narrative and car review
- Segmentation into atomic assertions
- Restriction to first 20 assertions of narrative and first 19 assertions of car review
- Presentation of title, author and non-declarative utterances (i.e. interrogatives and imperatives from car review) for completeness

You will read the beginning of **What Mrs Simpson Knows About Immigrants** by Kinneson Lalor.

Participants

- 61 speakers of British English between ages of 18 and 40 recruited on *Prolific*
- Presumably naïve with respect to QUDs
- 30 participants for narrative text (compensation of £2.60)
- 31 participants for longer car review (compensation of £3.50)

Data collection through online production experiment

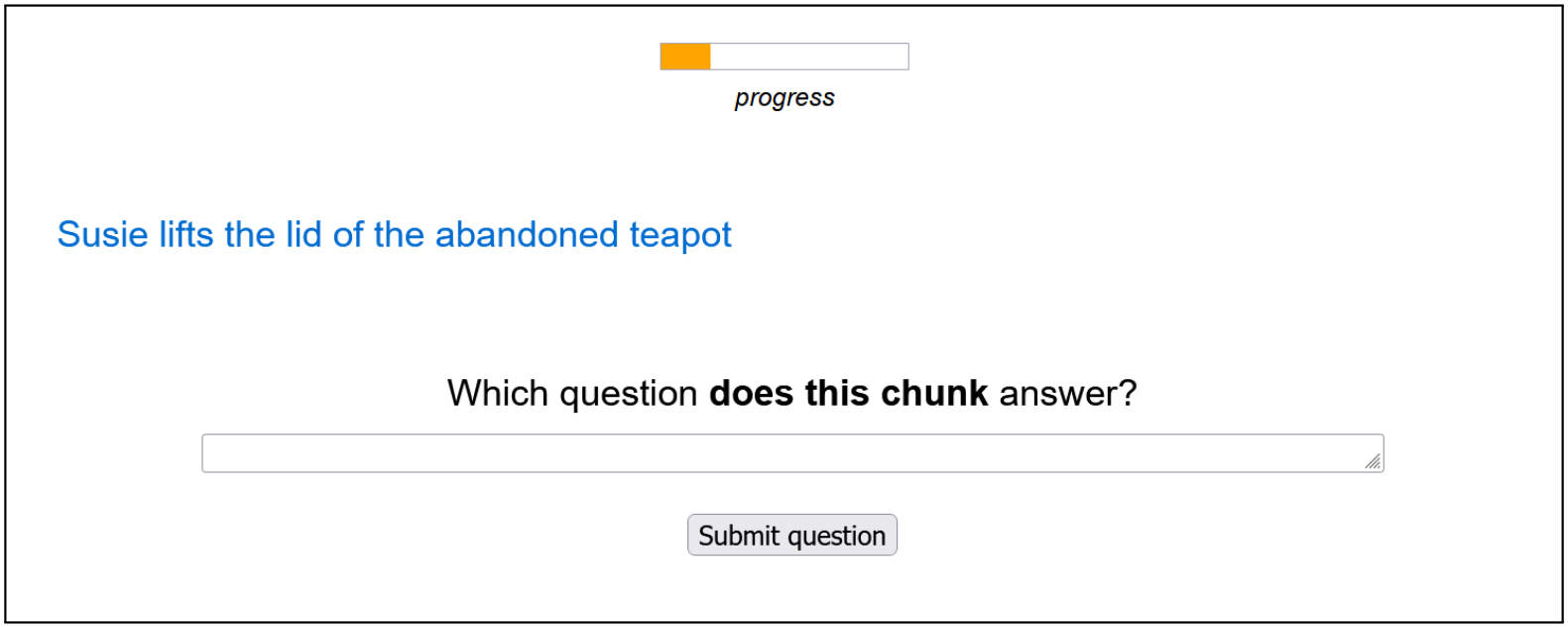
Procedure

- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers

Data collection through online production experiment

Procedure

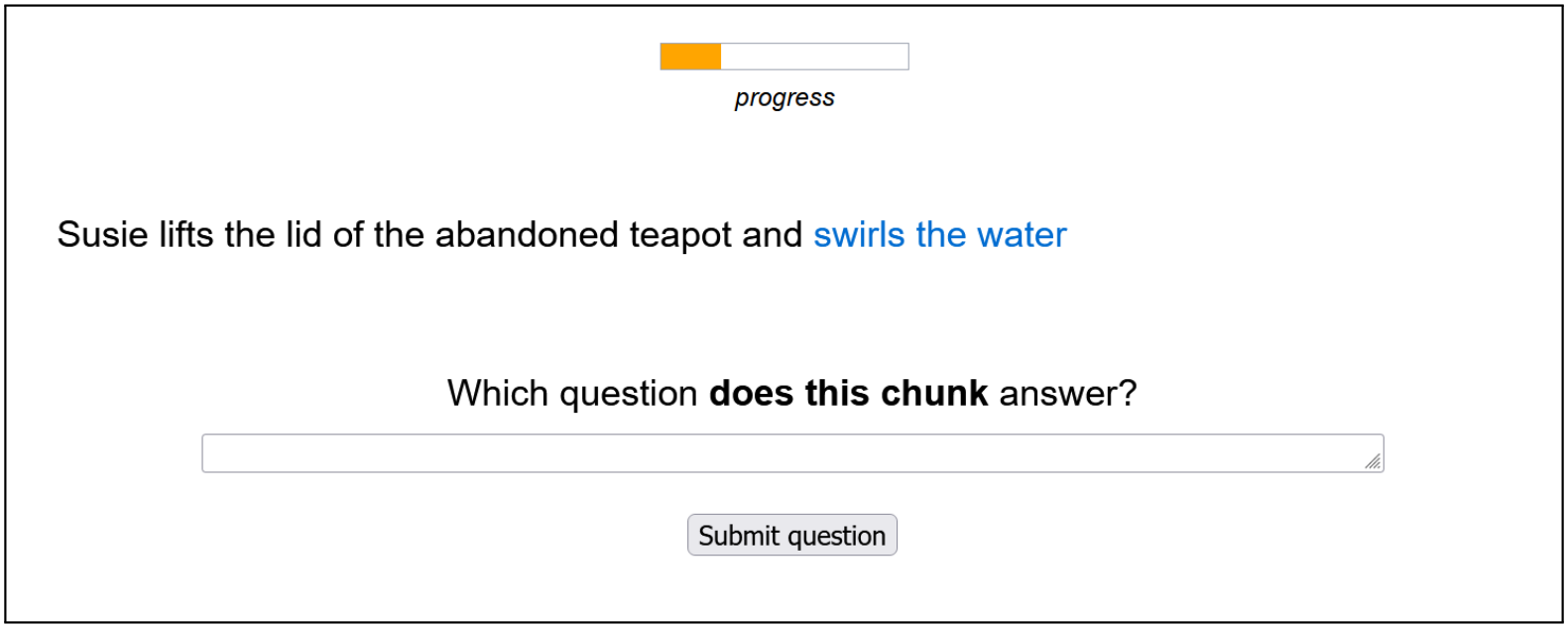
- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers



Data collection through online production experiment

Procedure

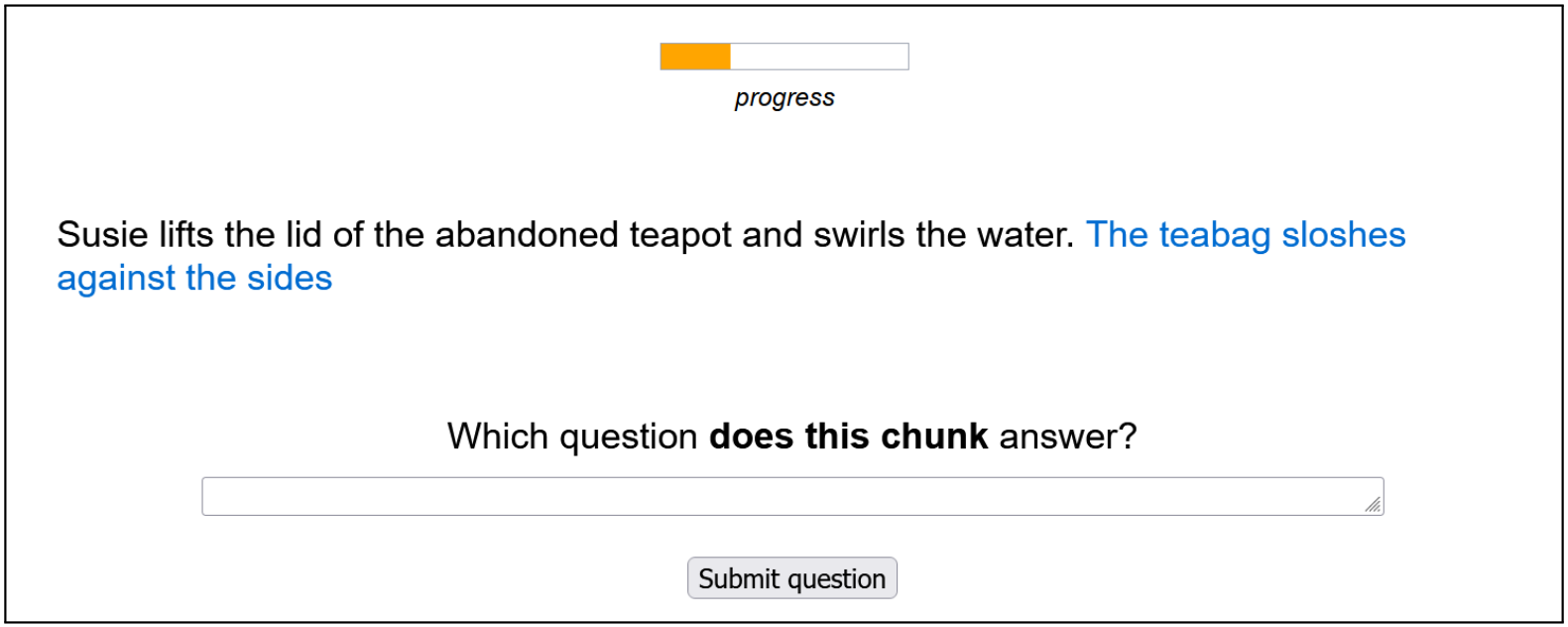
- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers



Data collection through online production experiment

Procedure

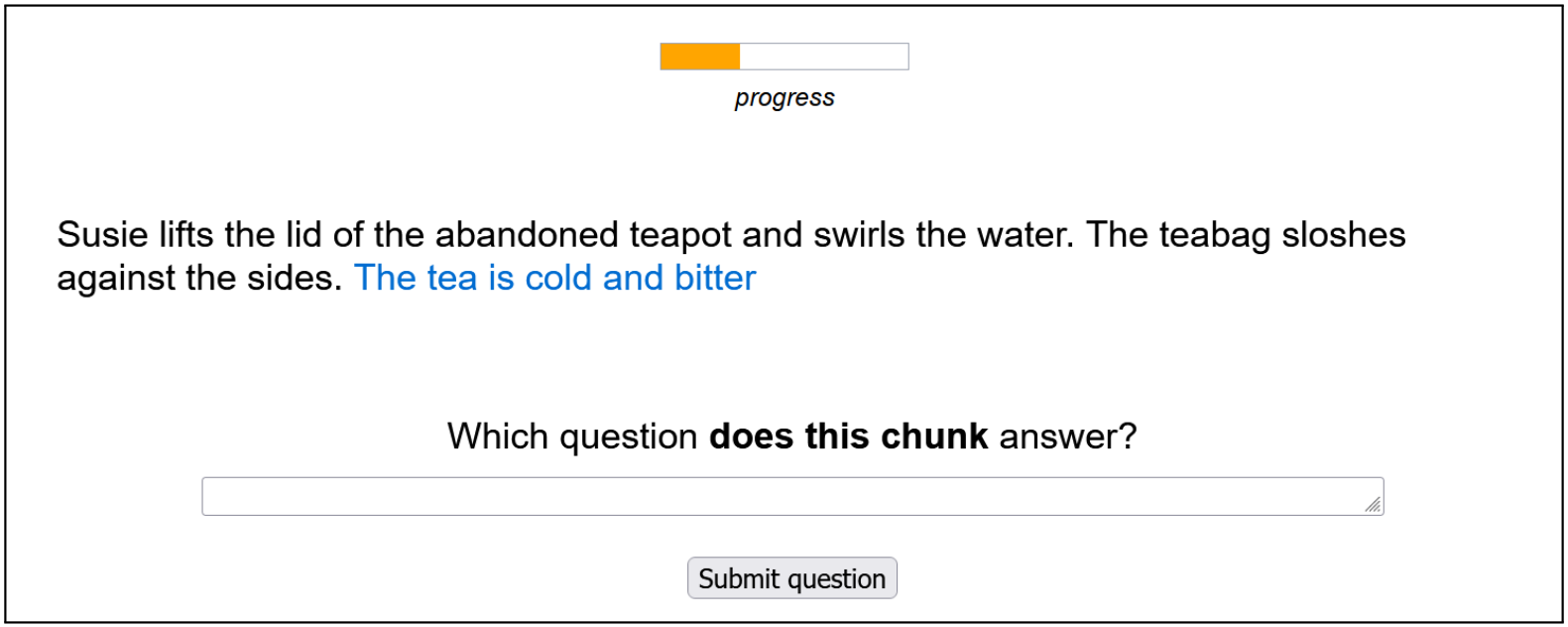
- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers



Data collection through online production experiment

Procedure

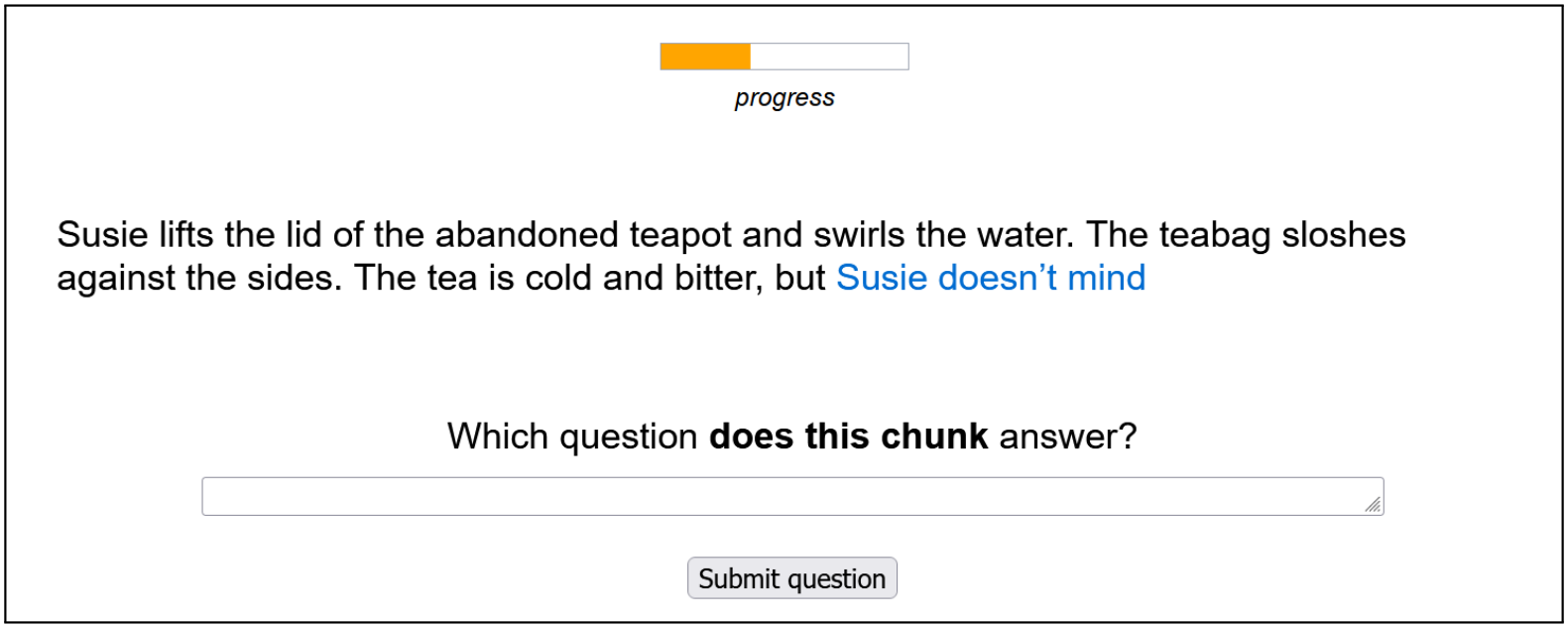
- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers



Data collection through online production experiment

Procedure

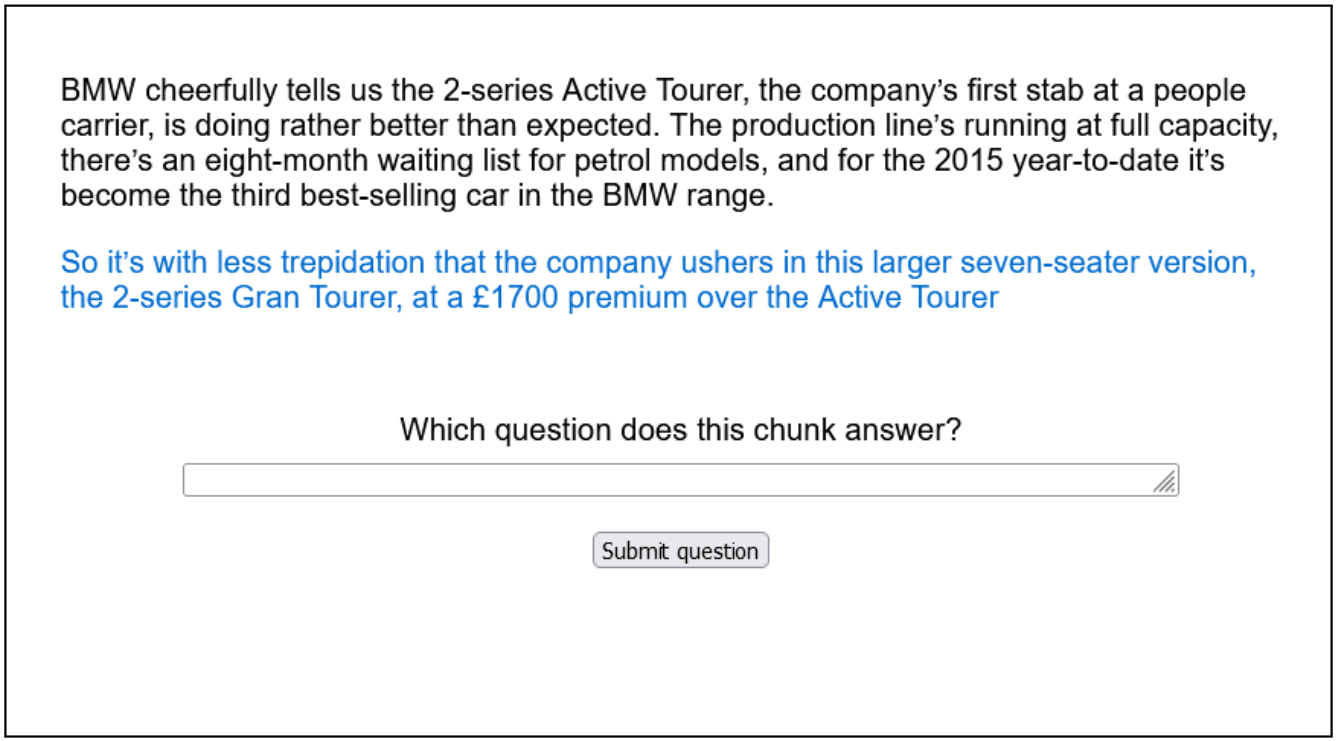
- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers



Data collection through online production experiment

Procedure

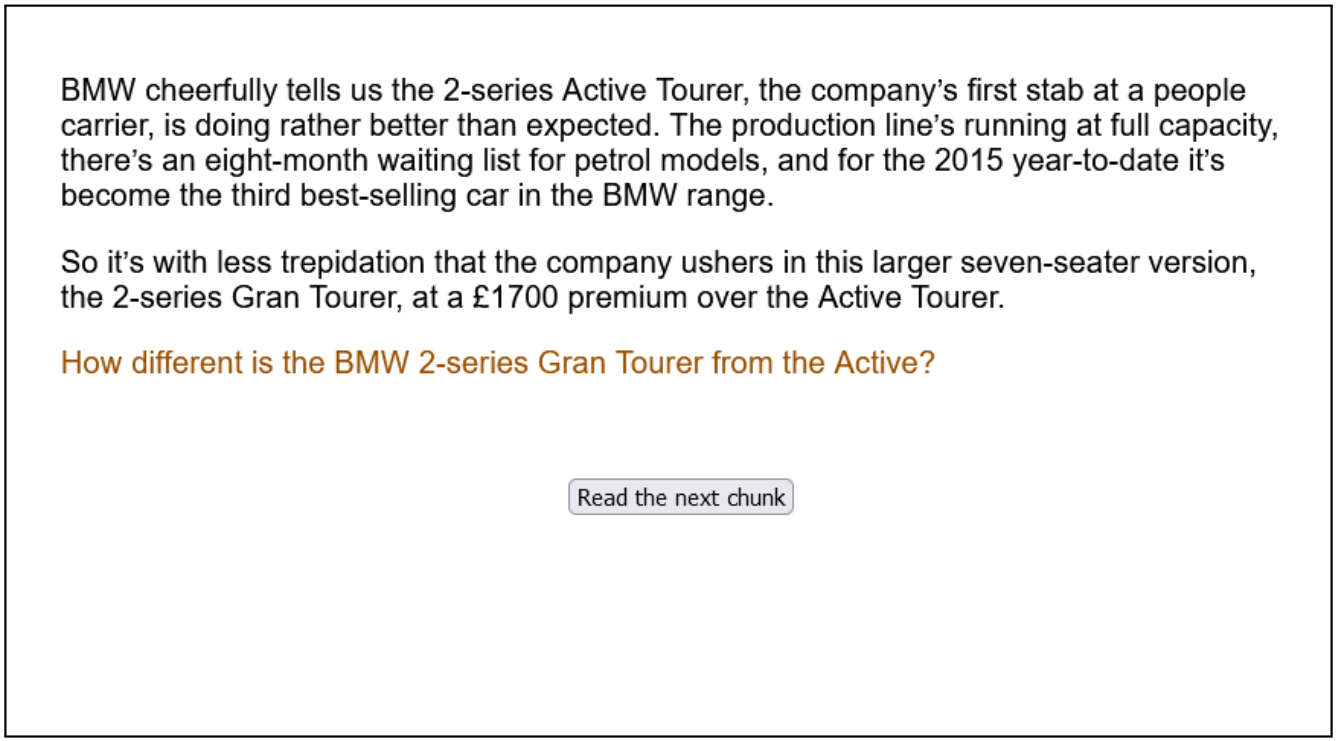
- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers



Data collection through online production experiment

Procedure

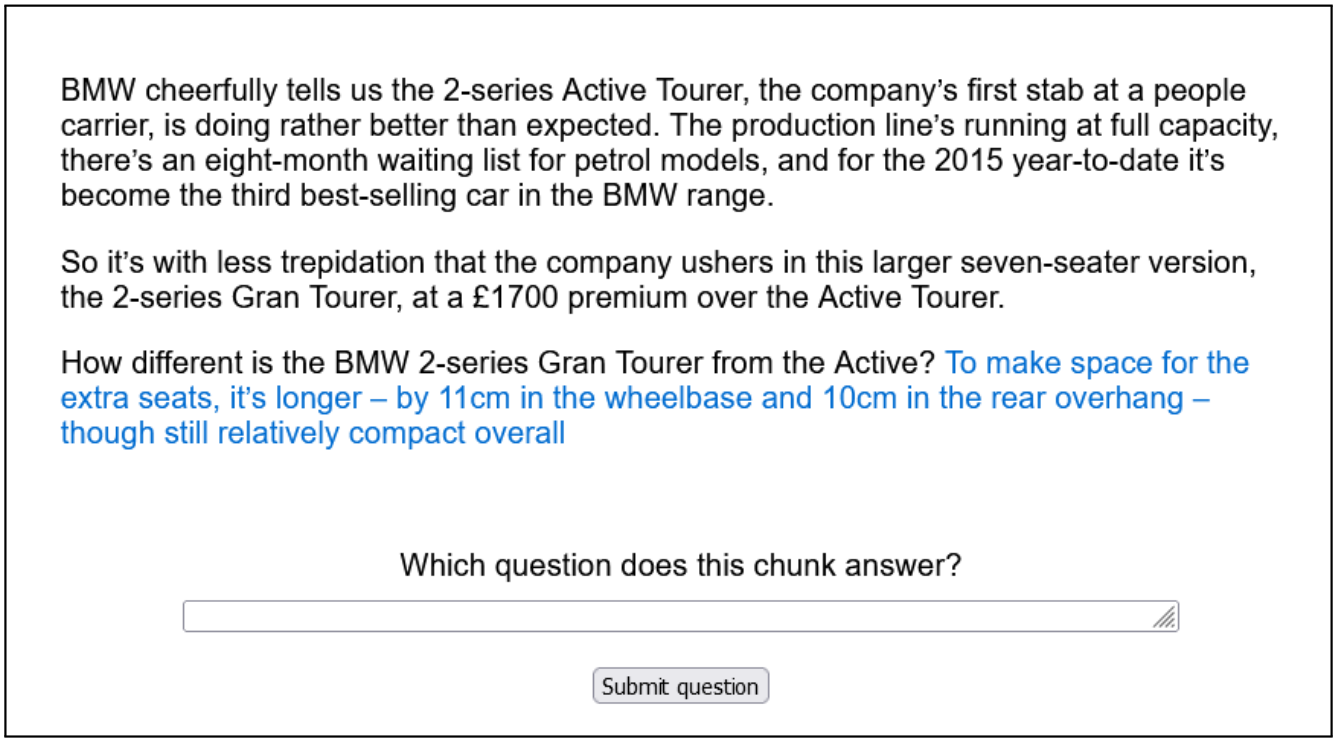
- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers



Data collection through online production experiment

Procedure

- Implementation of one survey per text with *PCIbex* (Zehr and Schwarz, 2018)
- Texts were presented assertion by assertion (precontext stayed visible)
- Assertions followed by a text field to enter the question the assertion answers



Data collection through online production experiment

Instructions

- Introduce participants informally to concept of QUDs:

Data collection through online production experiment

Instructions

- Introduce participants informally to concept of QUDs:
  - Declarative utterances as answers to potentially implicit question

Data collection through online production experiment

Instructions

- Introduce participants informally to concept of QUDs:
  - Declarative utterances as answers to potentially implicit question
  - One utterance can answer different questions

Data collection through online production experiment

Instructions

- Introduce participants informally to concept of QUDs:
  - Declarative utterances as answers to potentially implicit question
  - One utterance can answer different questions
  - Also parts of utterances can answer a question

Data collection through online production experiment

Instructions

- Introduce participants informally to concept of QUDs:
  - Declarative utterances as answers to potentially implicit question
  - One utterance can answer different questions
  - Also parts of utterances can answer a question
- Illustration with non-related examples:

Data collection through online production experiment

Instructions

- Introduce participants informally to concept of QUDs:
  - Declarative utterances as answers to potentially implicit question
  - One utterance can answer different questions
  - Also parts of utterances can answer a question
- Illustration with non-related examples:

(2) Mary and Ann went to an Italian restaurant.

(3) a. *What happened?*
b. *Where did Mary and Ann go?*

Data collection through online production experiment

Instructions

- Introduce participants informally to concept of QUDs:
  - Declarative utterances as answers to potentially implicit question
  - One utterance can answer different questions
  - Also parts of utterances can answer a question
- Illustration with non-related examples:

(2) Mary and Ann went to an Italian restaurant.

(3) a. *What happened?*
b. *Where did Mary and Ann go?*

- Participants should enter only one question per assertion
- Participants should enter most likely question
- Participants should not be funny / too creative

Data set creation

Preprocessing and exclusions

Preprocessing and exclusions

Excluded

Preprocessing and exclusions

Excluded

- 6.02% of the data (5.8% in the narrative text, 6.23% in the car review)

Preprocessing and exclusions

Excluded

- 6.02% of the data (5.8% in the narrative text, 6.23% in the car review)
- Questions related to the task rather than the text (e.g., *What have I missed?*)

Preprocessing and exclusions

Excluded

- 6.02% of the data (5.8% in the narrative text, 6.23% in the car review)
- Questions related to the task rather than the text (e.g., *What have I missed?*)
- Non-interrogative statements or bare DPs (e.g., *Tell me some details about Susie's life* or *waiting time*)

Preprocessing and exclusions

Excluded

- 6.02% of the data (5.8% in the narrative text, 6.23% in the car review)
- Questions related to the task rather than the text (e.g., *What have I missed?*)
- Non-interrogative statements or bare DPs (e.g., *Tell me some details about Susie's life* or *waiting time*)
- Parts of the utterance copied into the text field

Preprocessing and exclusions

Excluded

- 6.02% of the data (5.8% in the narrative text, 6.23% in the car review)
- Questions related to the task rather than the text (e.g., *What have I missed?*)
- Non-interrogative statements or bare DPs (e.g., *Tell me some details about Susie's life* or *waiting time*)
- Parts of the utterance copied into the text field

Not excluded

Preprocessing and exclusions

Excluded

- 6.02% of the data (5.8% in the narrative text, 6.23% in the car review)
- Questions related to the task rather than the text (e.g., *What have I missed?*)
- Non-interrogative statements or bare DPs (e.g., *Tell me some details about Susie's life* or *waiting time*)
- Parts of the utterance copied into the text field

Not excluded

- Subordinate questions lacking a matrix clause (e.g., *How the tea was*)

Preprocessing and exclusions

Excluded

- 6.02% of the data (5.8% in the narrative text, 6.23% in the car review)
- Questions related to the task rather than the text (e.g., *What have I missed?*)
- Non-interrogative statements or bare DPs (e.g., *Tell me some details about Susie's life* or *waiting time*)
- Parts of the utterance copied into the text field

Not excluded

- Subordinate questions lacking a matrix clause (e.g., *How the tea was*)
- More than one question produced for a single assertion by a single participant (e.g., *What is Susie doing? And how does she feel?*)

$\Rightarrow$ We entered each of the questions separately into the data set

Preprocessing and exclusions

Final data set

- Narrative text: 568 questions for 20 assertions
- Car review: 557 questions for 19 assertions

Annotation

Annotation

Semantic pooling of QUDs

Annotation

Semantic pooling of QUDs

Procedure

Assigning a unique single label to all semantically identical QUDs produced for a single assertion

Annotation

Semantic pooling of QUDs

Procedure

Assigning a unique single label to all semantically identical QUDs produced for a single assertion

Goal

Avoiding that the probability mass of a single QUD is split between synonymous expressions

Annotation

Semantic pooling of QUDs

Annotation

Semantic pooling of QUDs

- **Semantic identity:** Having the same set of answer propositions

(Hamblin, 1973; Karttunen, 1977)

Annotation

Semantic pooling of QUDs

- **Semantic identity:** Having the same set of answer propositions

(Hamblin, 1973; Karttunen, 1977)

Produced QUD	Label
What does Susie do?	*What does Susie do?*
What is Susie doing?	*What does Susie do?*
What did Susie do to the teapot?	*What did Susie do to the teapot?*
What does Susie do?	*What does Susie do?*

Example of a label assignment to QUDs produced for the utterance *Susie lifts the lid of the abandoned teapot*

Annotation

Semantic pooling of QUDs

- **Semantic identity:** Having the same set of answer propositions

(Hamblin, 1973; Karttunen, 1977)

Produced QUD	Label
What does Susie do?	*What does Susie do?*
What is Susie doing?	*What does Susie do?*
What did Susie do to the teapot?	*What did Susie do to the teapot?*
What does Susie do?	*What does Susie do?*

Example of a label assignment to QUDs produced for the utterance *Susie lifts the lid of the abandoned teapot*

- Label: The most frequent lexical realization of a QUD, corrected for spelling

Annotation

Semantic pooling of QUDs

- **Semantic identity:** Having the same set of answer propositions

(Hamblin, 1973; Karttunen, 1977)

Produced QUD	Label
What does Susie do?	*What does Susie do?*
What is Susie doing?	*What does Susie do?*
What did Susie do to the teapot?	*What did Susie do to the teapot?*
What does Susie do?	*What does Susie do?*

Example of a label assignment to QUDs produced for the utterance *Susie lifts the lid of the abandoned teapot*

- Label: The most frequent lexical realization of a QUD, corrected for spelling
- Single gold standard agreed upon by two expert annotators (Schäfer and Hristova)

Annotation

Semantic pooling of QUDs

- **Semantic identity:** Having the same set of answer propositions

(Hamblin, 1973; Karttunen, 1977)

Produced QUD	Label
What does Susie do?	*What does Susie do?*
What is Susie doing?	*What does Susie do?*
What did Susie do to the teapot?	*What did Susie do to the teapot?*
What does Susie do?	*What does Susie do?*

Example of a label assignment to QUDs produced for the utterance *Susie lifts the lid of the abandoned teapot*

- Label: The most frequent lexical realization of a QUD, corrected for spelling
- Single gold standard agreed upon by two expert annotators (Schäfer and Hristova)
- **Result:** A set of QUDs for each assertion

Challenges

Challenges

Calculating inter-annotator agreement for pooled data

Challenges

Calculating inter-annotator agreement for pooled data

- No fixed number of QUD labels per assertion

Challenges

Calculating inter-annotator agreement for pooled data

- No fixed number of QUD labels per assertion
- No fixed label names

Challenges

Calculating inter-annotator agreement for pooled data

- No fixed number of QUD labels per assertion
- No fixed label names

$\Rightarrow$ Not possible to use standard measures of inter-annotator agreement

Challenges

Calculating inter-annotator agreement for pooled data

- No fixed number of QUD labels per assertion
- No fixed label names

$\Rightarrow$ Not possible to use standard measures of inter-annotator agreement

Pragmatic information

- How much pragmatic information should we take into account when pooling QUDs?

Challenges

Calculating inter-annotator agreement for pooled data

- No fixed number of QUD labels per assertion
- No fixed label names

⇒ Not possible to use standard measures of inter-annotator agreement

Pragmatic information

- How much pragmatic information should we take into account when pooling QUDs?

(4) [*Susie lifts the lid of the abandoned teapot*] and [*swirls the water*].

(5) a. *What did Susie do?*
b. *What else did Susie do?*
c. *What did Susie do next?*

Challenges

- Some observations

Challenges

- Some observations
  - QUDs with *else* – more frequent for the non-initial conjuncts of coordinations than for other assertions ($\chi^2(1) = 4.17, p < 0.05$)

Challenges

- Some observations
  - QUDs with *else* – more frequent for the non-initial conjuncts of coordinations than for other assertions ($\chi^2(1) = 4.17, p < 0.05$)
  - QUDs with *next* – more frequent for the non-initial conjuncts of coordinations and utterances at the beginning of a paragraph than for other assertions ($\chi^2(1) = 4.26, p < 0.05$)

Challenges

- Some observations
  - QUDs with *else* – more frequent for the non-initial conjuncts of coordinations than for other assertions ($\chi^2(1) = 4.17, p < 0.05$)
  - QUDs with *next* – more frequent for the non-initial conjuncts of coordinations and utterances at the beginning of a paragraph than for other assertions ($\chi^2(1) = 4.26, p < 0.05$)

  $\Rightarrow$ QUDs containing *else* or *next* should not be grouped together with QUDs lacking this linguistic material

Challenges

- Some observations
  - QUDs with *else* – more frequent for the non-initial conjuncts of coordinations than for other assertions ($\chi^2(1) = 4.17, p < 0.05$)
  - QUDs with *next* – more frequent for the non-initial conjuncts of coordinations and utterances at the beginning of a paragraph than for other assertions ($\chi^2(1) = 4.26, p < 0.05$)

  $\Rightarrow$ QUDs containing *else* or *next* should not be grouped together with QUDs lacking this linguistic material

Back to the question at hand

- How much pragmatic information should we take into account when pooling QUDs?

Challenges

- Some observations
  - QUDs with *else* – more frequent for the non-initial conjuncts of coordinations than for other assertions ($\chi^2(1) = 4.17, p < 0.05$)
  - QUDs with *next* – more frequent for the non-initial conjuncts of coordinations and utterances at the beginning of a paragraph than for other assertions ($\chi^2(1) = 4.26, p < 0.05$)

$\Rightarrow$ QUDs containing *else* or *next* should not be grouped together with QUDs lacking this linguistic material

Back to the question at hand

- How much pragmatic information should we take into account when pooling QUDs?

$\Rightarrow$ We chose a purely semantic approach, where only the propositional content of the QUD was considered

Structure of the data set

Structure of the data set

Structure of the data set

For each **QUD** from a set of possible QUDs, we calculated:

Structure of the data set

For each **QUD** from a set of possible QUDs, we calculated:

- Its **frequency rank** with respect to the other QUDs within its set

Structure of the data set

For each **QUD** from a set of possible QUDs, we calculated:

- Its **frequency rank** with respect to the other QUDs within its set
- Its **probability** with respect to the other QUDs within its set (cf. equation 1)

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

Structure of the data set

For each **QUD** from a set of possible QUDs, we calculated:

- Its **frequency rank** with respect to the other QUDs within its set
- Its **probability** with respect to the other QUDs within its set (cf. equation 1)

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

QUD	n	Rank	Probability
What did Susie do?	11	1	0.37
What did Susie do to the teapot?	6	2	0.2
What did Susie lift?	4	3	0.13
What did Susie lift the lid of?	2	4	0.07
What did Susie do next?	1	5	0.03
…	…	…	…

Section of the probability distribution for the first assertion of the narrative text

Structure of the data set

For each **assertion**, we calculated the **entropy H** in the probability distribution over QUDs in its QUD set (cf. equation 2)

$$H = -\sum_{i=1}^{n} p_i log_2 p_i \tag{2}$$

Structure of the data set

For each **assertion**, we calculated the **entropy H** in the probability distribution over QUDs in its QUD set (cf. equation 2)

$$H = -\sum_{i=1}^{n} p_i log_2 p_i \tag{2}$$

- Measures the degree of uncertainty about the outcome of a random variable

(Shannon, 1948, p. 393)

Structure of the data set

For each **assertion**, we calculated the **entropy H** in the probability distribution over QUDs in its QUD set (cf. equation 2)

$$H = -\sum_{i=1}^{n} p_i log_2 p_i \tag{2}$$

- Measures the degree of uncertainty about the outcome of a random variable (Shannon, 1948, p. 393)
- Maximal if all of the QUDs in a set are equally likely

Structure of the data set

For each **assertion**, we calculated the **entropy H** in the probability distribution over QUDs in its QUD set (cf. equation 2)

$$H = -\sum_{i=1}^{n} p_i log_2 p_i \tag{2}$$

- Measures the degree of uncertainty about the outcome of a random variable (Shannon, 1948, p. 393)
- Maximal if all of the QUDs in a set are equally likely
- Equals 0 if there is only one QUD

Structure of the data set

For each **assertion**, we calculated the **entropy H** in the probability distribution over QUDs in its QUD set (cf. equation 2)

$$H = -\sum_{i=1}^{n} p_i log_2 p_i \tag{2}$$

- Measures the degree of uncertainty about the outcome of a random variable (Shannon, 1948, p. 393)
- Maximal if all of the QUDs in a set are equally likely
- Equals 0 if there is only one QUD

QUD	Probability
QUD_1	0.25
QUD_2	0.25
QUD_3	0.25
QUD_4	0.25

QUD	Probability
QUD_1	1

Examples of a probability distribution over QUDs with high (left) and low (right) entropy

XML schema

XML schema

The QUD element

XML schema

The QUD element

- Represents a QUD from a set of possible QUDs for an assertion

XML schema

The QUD element

- Represents a QUD from a set of possible QUDs for an assertion
- Contains a UNIT element representing the assertion which answers the QUD

XML schema

The QUD element

- Represents a QUD from a set of possible QUDs for an assertion
- Contains a UNIT element representing the assertion which answers the QUD

$\Rightarrow$ Each assertion occurs several times in the data set

XML schema

The QUD element

- Represents a QUD from a set of possible QUDs for an assertion
- Contains a UNIT element representing the assertion which answers the QUD

⇒ Each assertion occurs several times in the data set

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

The XML structure for the beginning of the narrative text

Data set statistics

Variation between QUDs

- Most of the time, one clearly preferred QUD
- QUD on rank 1: mean p = .28 (sd=.12), QUD on rank 2: mean p = .13 (sd=.06)
- In both texts, no sentence with less than 8 different QUDs
- This holds in spite of subjects producing **only the most likely** QUD

Example: Skewed distribution

(6) The skin feels puckered and soft, like white and wrinkly fingertips in the bath

QUD	*n*	Rank	Probability
How does the skin feel?	18	1	0.62
How did the cap from work affect Susie's skin?	1	2	0.03
How does Susie's head feel?	1	2	0.03
Then what happened?	1	2	0.03
Why is Susie rubbing her head?	1	2	0.03
…	…	…	…

Example: (Relatively) flat distribution

(7) Range-topper is the 2.0-litre, four-cylinder 187bhp 220d xDrive, capable of dipping under eight seconds from 0-62mph

QUD	*n*	Rank	Probability
How fast is it?	2	1	0.07
What's the top acceleration?	2	1	0.07
What are the engine specifications for best range in this model?	2	1	0.07
What are the engine specs for the top model?	2	1	0.07
What engine does the range topper have?	2	1	0.07
What are the performance figures?	1	6	0.03
…	…	…	…

Differences between text types

- Higher likelihood for QUD on rank 1 in narrative (.34) than in car review (.21)
- Higher mean number of QUDs in car review (18.58) than in narrative (14.6)
- Higher entropy in car review (3.81) than in narrative (3.22)

Differences between text types

- Higher likelihood for QUD on rank 1 in narrative (.34) than in car review (.21)
- Higher mean number of QUDs in car review (18.58) than in narrative (14.6)
- Higher entropy in car review (3.81) than in narrative (3.22)

Potential explanations

- Mean complexity of assertion (n words) $\rightarrow$ More potential QUDs

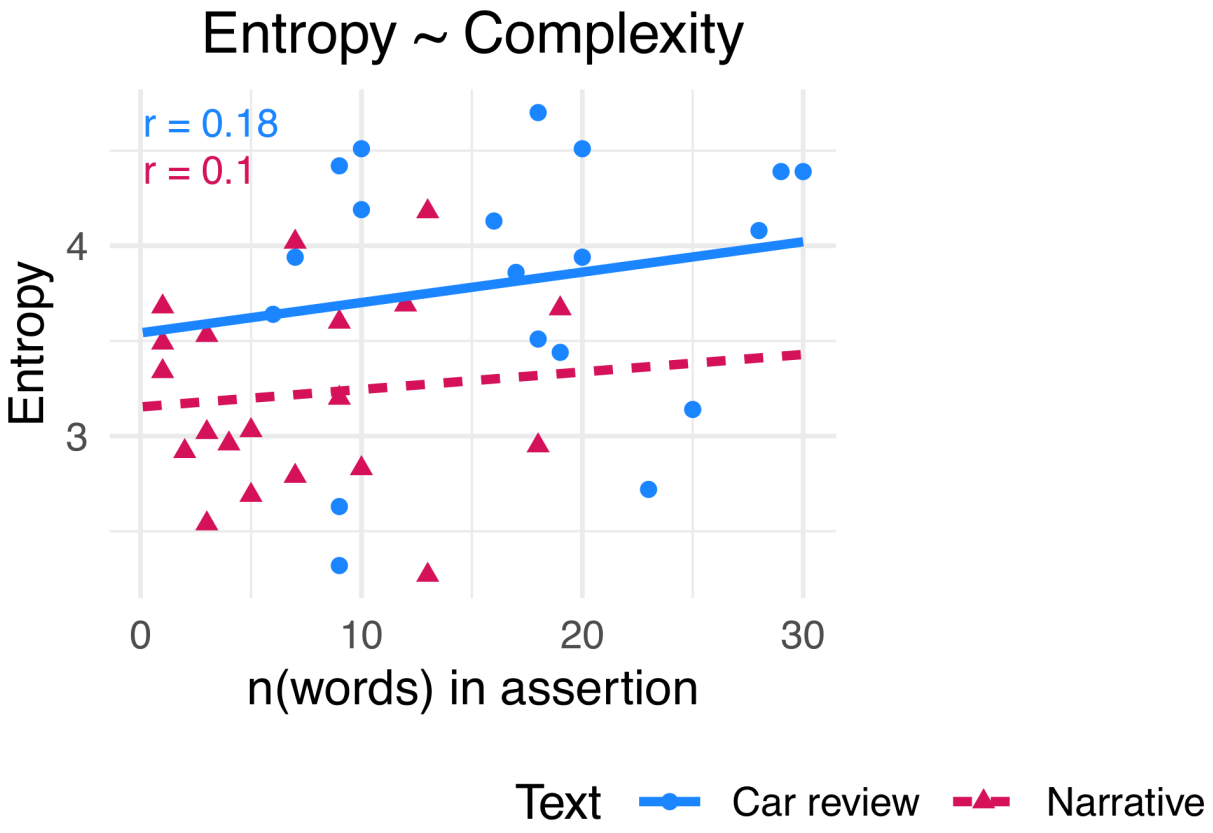
Differences between text types

- Higher likelihood for QUD on rank 1 in narrative (.34) than in car review (.21)
- Higher mean number of QUDs in car review (18.58) than in narrative (14.6)
- Higher entropy in car review (3.81) than in narrative (3.22)

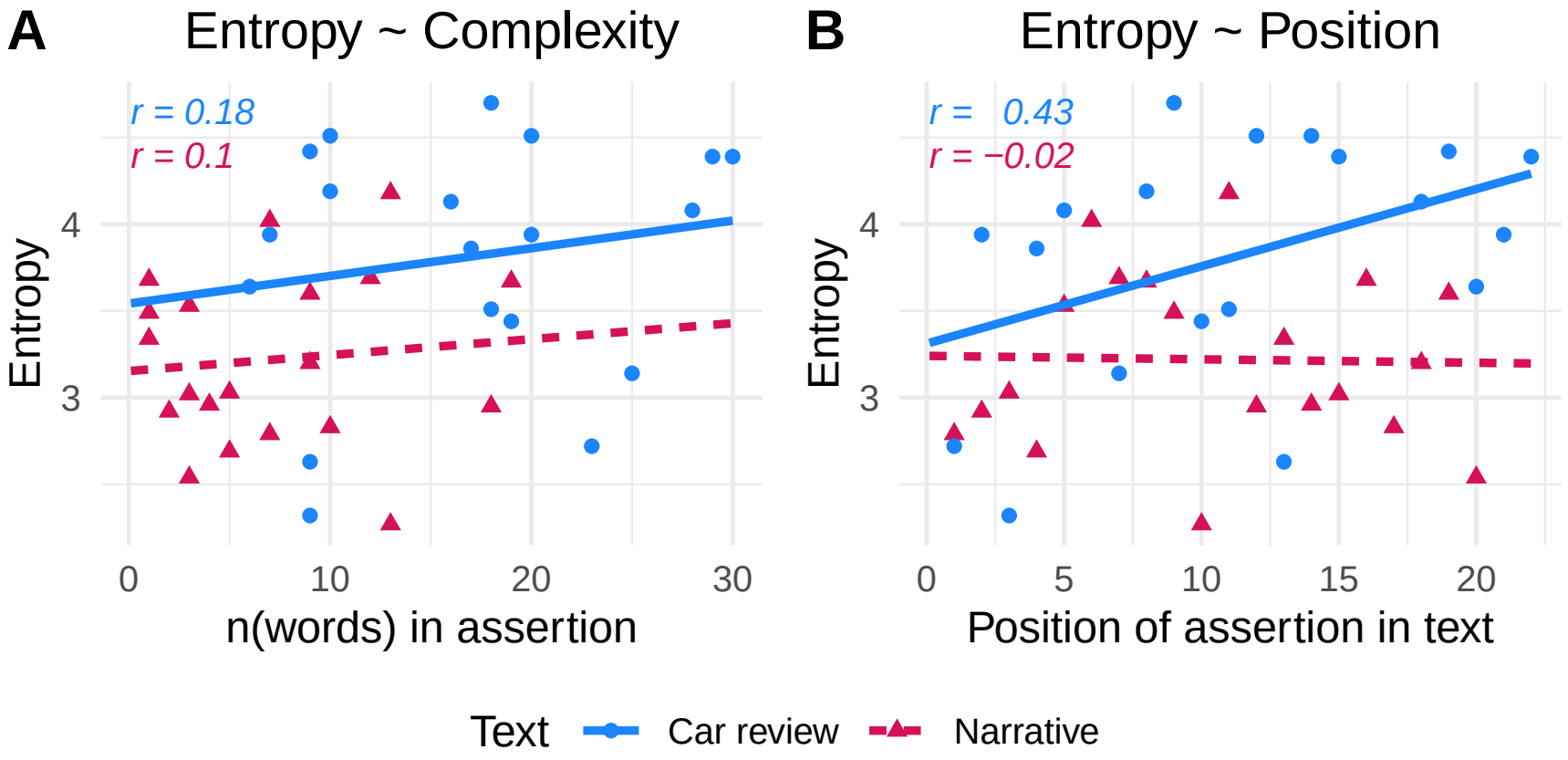
Potential explanations

- Mean complexity of assertion (n words) $\rightarrow$ More potential QUDs
- Position of assertion in text
  - The later, the less entropy: Topic is narrowed down through discourse
  - The later, the more entropy: More potential topics later in discourse

Distribution of entropy across the text



Distribution of entropy across the text



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

Wrapping up

Wrapping up

Quantitative approach to QUD annotation

- Often, one QUD is clearly the most likely
- Considerable amount of variation among the other QUDs

$\Rightarrow$ Need for quantitative model of QUD-based discourse structure?

Wrapping up

Quantitative approach to QUD annotation

- Often, one QUD is clearly the most likely
- Considerable amount of variation among the other QUDs

$\Rightarrow$ Need for quantitative model of QUD-based discourse structure?

Open questions

Wrapping up

Quantitative approach to QUD annotation

- Often, one QUD is clearly the most likely
- Considerable amount of variation among the other QUDs

$\Rightarrow$ Need for quantitative model of QUD-based discourse structure?

Open questions

- To what extent are our data in line with the (more fine-grained) expert annotations? (e.g. in terms of the most often produced QUD)

Wrapping up

Quantitative approach to QUD annotation

- Often, one QUD is clearly the most likely
- Considerable amount of variation among the other QUDs

$\Rightarrow$ Need for quantitative model of QUD-based discourse structure?

Open questions

- To what extent are our data in line with the (more fine-grained) expert annotations? (e.g. in terms of the most often produced QUD)
- Which factors determine the entropy in each utterance's QUD set?

Wrapping up

Quantitative approach to QUD annotation

- Often, one QUD is clearly the most likely
- Considerable amount of variation among the other QUDs

$\Rightarrow$ Need for quantitative model of QUD-based discourse structure?

Open questions

- To what extent are our data in line with the (more fine-grained) expert annotations? (e.g. in terms of the most often produced QUD)
- Which factors determine the entropy in each utterance's QUD set?
- Are the focus-background structure of utterances and QUDs aligned?

Wrapping up

Quantitative approach to QUD annotation

- Often, one QUD is clearly the most likely
- Considerable amount of variation among the other QUDs

$\Rightarrow$ Need for quantitative model of QUD-based discourse structure?

Open questions

- To what extent are our data in line with the (more fine-grained) expert annotations? (e.g. in terms of the most often produced QUD)
- Which factors determine the entropy in each utterance's QUD set?
- Are the focus-background structure of utterances and QUDs aligned?
- Do the QUDs produced address NAI content?

(8) her landlady, Mrs Simpson, normally reuses tea bags
QUD: "Who is Mrs Simpson?"

QUD likelihood in context or given an assertion?

Our approach: $p(QUD|assertion, context)$

- QUD is reconstructed given the corresponding assertion after seeing it

QUD likelihood in context or given an assertion?

Our approach: $p(QUD|assertion, context)$

- QUD is reconstructed given the corresponding assertion after seeing it

Discourse-based expectation about QUDs: $p(QUD|context)$

- QUDs might be raised by preceding material
- Alternative task: Guess which question the **next** sentence in text answers

QUD likelihood in context or given an assertion?

Our approach: $p(QUD|assertion, context)$

- QUD is reconstructed given the corresponding assertion after seeing it

Discourse-based expectation about QUDs: $p(QUD|context)$

- QUDs might be raised by preceding material
- Alternative task: Guess which question the **next** sentence in text answers
- Pilot study (only some UdS colleagues and student assistants)
  - High variation between QUDs, participants report not "getting it right" often (9)

(9) Usually, by the time Susie gets home, the tea tastes mostly of chlorine.
[QUD: When does Susie get home (usually)?]
As she checks Mrs Simpson's calendar, Susie rubs the place where the elastic cap from work scrunched all day.

QUD likelihood in context or given an assertion?

Our approach: $p(QUD|assertion, context)$

- QUD is reconstructed given the corresponding assertion after seeing it

Discourse-based expectation about QUDs: $p(QUD|context)$

- QUDs might be raised by preceding material
- Alternative task: Guess which question the **next** sentence in text answers
- Pilot study (only some UdS colleagues and student assistants)
  - High variation between QUDs, participants report not "getting it right" often (9)
  - Some overt connectives indicate upcoming QUD, or narrative continuation (10)

(9) Usually, by the time Susie gets home, the tea tastes mostly of chlorine.
[QUD: When does Susie get home (usually)?]
As she checks Mrs Simpson's calendar, Susie rubs the place where the elastic cap from work scrunched all day.

(10) And what happened next?

References

De Kuthy, Kordula, Nils Reiter, and Arndt Riester (2018). “QUD-based Annotation of Discourse Structure and Information Structure: Tool and Evaluation”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018. Miyazaki, Japan: European Language Resources Association (ELRA).

Hamblin, Charles L. (1973). “Questions in Montague English”. In: *Foundations of Language* 10.1, pp. 41–53. JSTOR: 25000703.

Karttunen, Lauri (1977). “Syntax and Semantics of Questions”. In: *Linguistics and Philosophy* 1.1, pp. 3–44. DOI: 10.1007/BF00351935.

Morgan, Jerry (1973). “Sentence Fragments and the Notion ‘sentence’”. In: *Issues in Linguistics. Papers in Honor of Henry and Renée Kahane*. Ed. by Braj B. Kachru et al. Urbana: University of Illionois Press, pp. 719–751.

Poppels, Till and Andrew Kehler (to appear). “Ellipsis and the QUD: Sluicing with Nominal Antecedents”. In: *Information Structure and Discourse in Generative Grammar: Mechanisms and Processes*. Ed. by Andreas Konietzko and Susanne Winkler. Berlin, Boston: De Gruyter Mouton.

Reich, Ingo, Robin Lemke, and Lisa Schäfer (to appear). “Questions under Discussion, Salience and the Acceptability of Fragments.”. In: *Information Structure and Discourse in Generative Grammar: Mechanisms and Processes*. Ed. by Andreas Konietzko and Susanne Winkler. Berlin, Boston: De Gruyter Mouton.

Riester, Arndt (2019). “Constructing QUD Trees”. In: *Questions in Discourse. Volume 2: Pragmatics*. Ed. by Malte Zimmermann, Klaus von Heusinger, and Edgar Onea. Current Research in the Semantics / Pragmatics Interface 36. Leiden: Brill, pp. 164–193. DOI: 10.1163/9789004378322_007.

References

Riester, Arndt, Lisa Brunetti, and Kordula De Kuthy (2018). “Annotation Guidelines for Questions under Discussion and Information Structure”. In: *Information Structure in Lesser-described Languages: Studies in Prosody and Syntax*. Ed. by Evangelia Adamou, Katharina Haude, and Martine Vanhove. Studies in Language Companion Series 199. Amsterdam, Philadelphia: John Benjamins Publishing Company, pp. 403–444. DOI: 10.1075/slcs.199.14rie.

Roberts, Craige (2012). “Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics”. In: *Semantics and Pragmatics* 5, pp. 1–69. DOI: 10.3765/sp.5.6.

Shannon, Claude E. (1948). “A Mathematical Theory of Communication”. In: *The Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Van Kuppevelt, Jan (1995). “Discourse Structure, Topicality and Questioning”. In: *Journal of Linguistics* 31.1, pp. 109–147. DOI: 10.1017/S002222670000058X.

Von Stutterheim, Christiane and Wolfgang Klein (1989). “Referential Movement in Descriptive and Narrative Discourse”. In: *Language Processing in Social Context*. Ed. by Rainer Dietrich and Carl F. Graumann. Amsterdam: North Holland, pp. 39–76.

Westera, Matthijs, Laia Mayol, and Hannah Rohde (2020). “TED-Q: TED Talks and the Questions They Evoke”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Marseille, France: European Language Resources Association, pp. 1118–1127.

Zehr, Jérémy and Florian Schwarz (2018). “PennController for Internet Based Experiments (IBEX)”. In: DOI: 10.17605/OSF.IO/MD832.