Studying the anticipation of QUDs and discourse relations by crowdsourcing a dataset of 'evoked questions'

Matthijs Westera, Laia Mayol, Hannah Rohde
(Edinburgh)

Studying the anticipation of QUDs and discourse relations by crowdsourcing a dataset of 'evoked questions'

Matthijs Westera, Laia Mayol, Hannah Rohde
(Edinburgh)

Evoked questions

[...] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Evoked questions

[...] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Why wouldn't they use their prostheses?

Evoked questions

[...] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Why wouldn't they use their prostheses?
Why do they not use their prostheses?

Evoked questions

[…] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Why wouldn't they use their prostheses?
Why do they not use their prostheses?
Did they do something to help amputees?

Evoked questions

[...] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Why wouldn't they use their prostheses?
Why do they not use their prostheses?
Did they do something to help amputees?
Why wouldnt they use their prostesis?

Evoked questions

[…] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Why wouldn't they use their prostheses?
Why do they not use their prostheses?
Did they do something to help amputees?
Why wouldnt they use their prostesis?
How old are you now?

Evoked questions

[...] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Why wouldn't they use their prostheses?
Why do they not use their prostheses?
Did they do something to help amputees?
Why wouldnt they use their prostesis?
How old are you now?
Why didn't amputees use their prostheses?

Research questions

Could evoked questions provide a useful empirical window on discourse structure?

Research questions

Could evoked questions provide a useful empirical window on discourse structure?

- In particular, can we use them to answer:

Research questions

Could evoked questions provide a useful empirical window on discourse structure?

- In particular, can we use them to answer:

Is discourse structure more explicitly signaled in places where it is less predictable?

Research questions

Could evoked questions provide a useful empirical window on discourse structure?

Why would they?

- In particular, can we use them to answer:

Is discourse structure more explicitly signaled in places where it is less predictable?

Questions Under Discussion

e.g. Roberts 2012

Questions Under Discussion

Central tenets of the QUD framework: e.g. Roberts 2012

Questions Under Discussion

Central tenets of the QUD framework: e.g. Roberts 2012

- Discourse is (can be modeled as) a process of raising and resolving questions.

Questions Under Discussion

Central tenets of the QUD framework: e.g. Roberts 2012

- Discourse is (can be modeled as) a process of raising and resolving questions. ≠interrogatives

Questions Under Discussion

Central tenets of the QUD framework: e.g. Roberts 2012

- Discourse is (can be modeled as) a process of raising and resolving questions. ≠interrogatives
- These questions can be implicit or made (partially) explicit.

Questions Under Discussion

Central tenets of the QUD framework: e.g. Roberts 2012

- Discourse is (can be modeled as) a process of raising and resolving questions. ≠interrogatives
- These questions can be implicit or made (partially) explicit.
- The QUD of a given discourse move is the question it targets.

Questions Under Discussion

Central tenets of the QUD framework: e.g. Roberts 2012

- Discourse is (can be modeled as) a process of raising and resolving questions. ≠interrogatives
- These questions can be implicit or made (partially) explicit.
- The QUD of a given discourse move is the question it targets.
- Discourse progresses from question to question in a reasonable/rational/natural way.

… and evoked questions?

… and evoked questions?

- If part of a discourse strongly *evokes* a certain question…

… and evoked questions?

- If part of a discourse strongly *evokes* a certain question…
- …and this question happens to be answered by the next discourse move…

… and evoked questions?

- If part of a discourse strongly *evokes* a certain question…
- …and this question happens to be answered by the next discourse move…
- …then that question is very likely its QUD.

The example again

[...] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Why wouldn't they use their prostheses?
Why do they not use their prostheses?
Did they do something to help amputees?
Why wouldnt they use their prostesis?
How old are you now?
Why didn't amputees use their prostheses?

The example again

[...] As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses.

Why wouldn't they use their prostheses?
Why do they not use their prostheses?
Did they do something to help amputees?
Why wouldnt they use their prostesis?
How old are you now?
Why didn't amputees use their prostheses?

The reason, I would come to find out, was that their prosthetic sockets were painful [...]

Existing discourse annotations

Existing discourse annotations

- Existing discourse annotations mostly use Discourse Relations (e.g., PDTB)

Existing discourse annotations

- Existing discourse annotations mostly use Discourse Relations (e.g., PDTB)
  - Annotated as connectives (*because*, *hence*, *but*, ...)

Existing discourse annotations

- Existing discourse annotations mostly use Discourse Relations (e.g., PDTB)
  - Annotated as connectives (*because*, *hence*, *but*, …)
  - Fixed, smallish taxonomy

Existing discourse annotations

- Existing discourse annotations mostly use Discourse Relations (e.g., PDTB)
  - Annotated as connectives (*because*, *hence*, *but*, …)
  - Fixed, smallish taxonomy
- No large-scale QUD-annotation (cf. Riester et al.)

Existing discourse annotations

- Existing discourse annotations mostly use Discourse Relations (e.g., PDTB)
  - Annotated as connectives (*because*, *hence*, *but*, ...)
  - Fixed, smallish taxonomy
- No large-scale QUD-annotation (cf. Riester et al.)
  - Annotated as full interrogative sentences;

Existing discourse annotations

- Existing discourse annotations mostly use Discourse Relations (e.g., PDTB)
  - Annotated as connectives (*because*, *hence*, *but*, …)
  - Fixed, smallish taxonomy
- No large-scale QUD-annotation (cf. Riester et al.)
  - Annotated as full interrogative sentences;
  - No fixed taxonomy;

Existing discourse annotations

- Existing discourse annotations mostly use Discourse Relations (e.g., PDTB)
  - Annotated as connectives (*because*, *hence*, *but*, ...)
  - Fixed, smallish taxonomy
- No large-scale QUD-annotation (cf. Riester et al.)
  - Annotated as full interrogative sentences;
  - No fixed taxonomy;
  - Somewhat unnatural task.

Existing discourse annotations

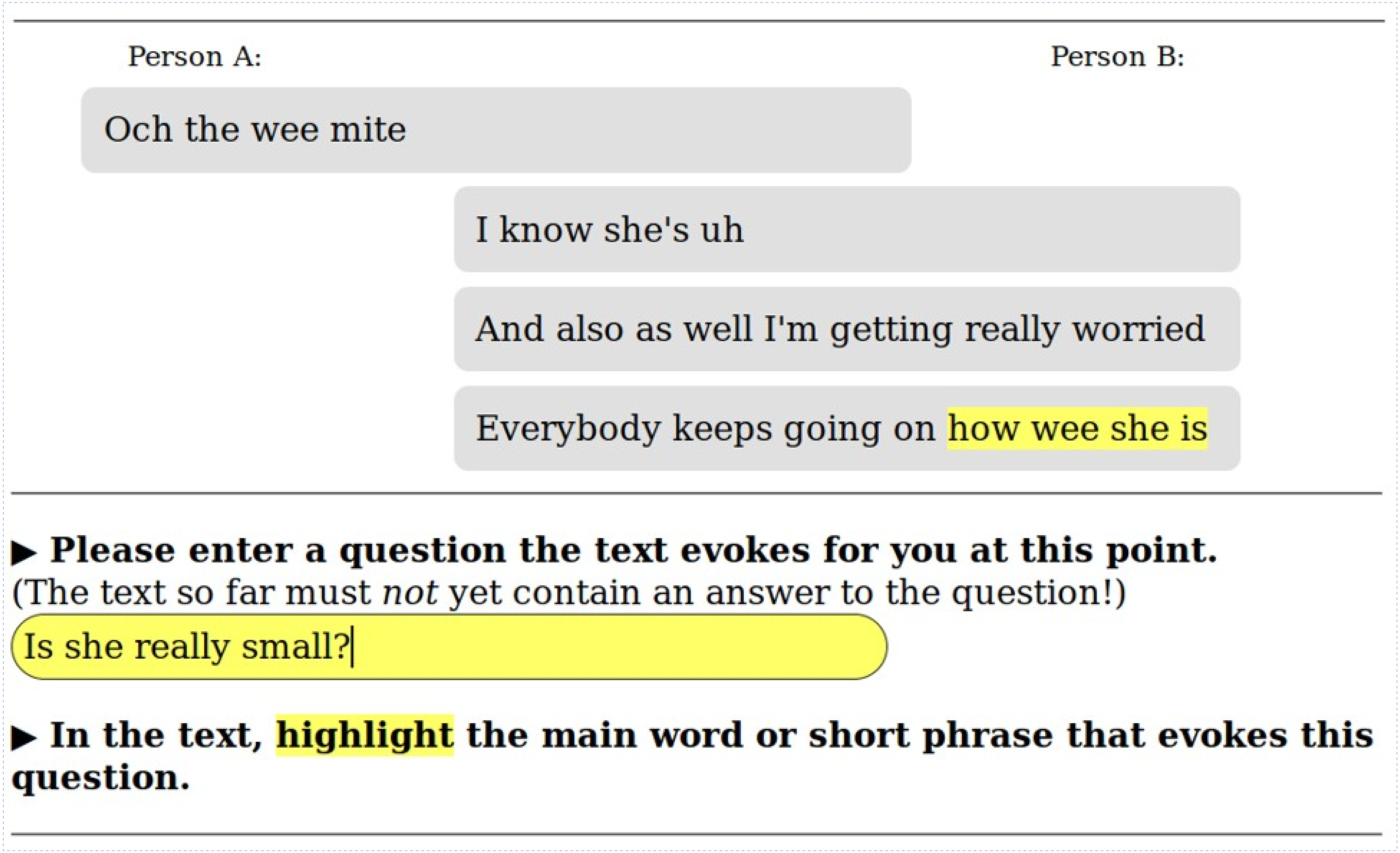
- Existing discourse annotations mostly use Discourse Relations (e.g., PDTB)
  - Annotated as connectives (*because*, *hence*, *but*, …)
  - Fixed, smallish taxonomy
- No large-scale QUD-annotation (cf. Riester et al.)
  - Annotated as full interrogative sentences;
  - No fixed taxonomy;
  - Somewhat unnatural task.

Unless we do it via evoked questions!

Elicitation tool

spellout.net/ibexexps/mwestera/evoque/

Elicitation tool



Person A:

Person B:

Och the wee mite

I know she's uh

And also as well I'm getting really worried

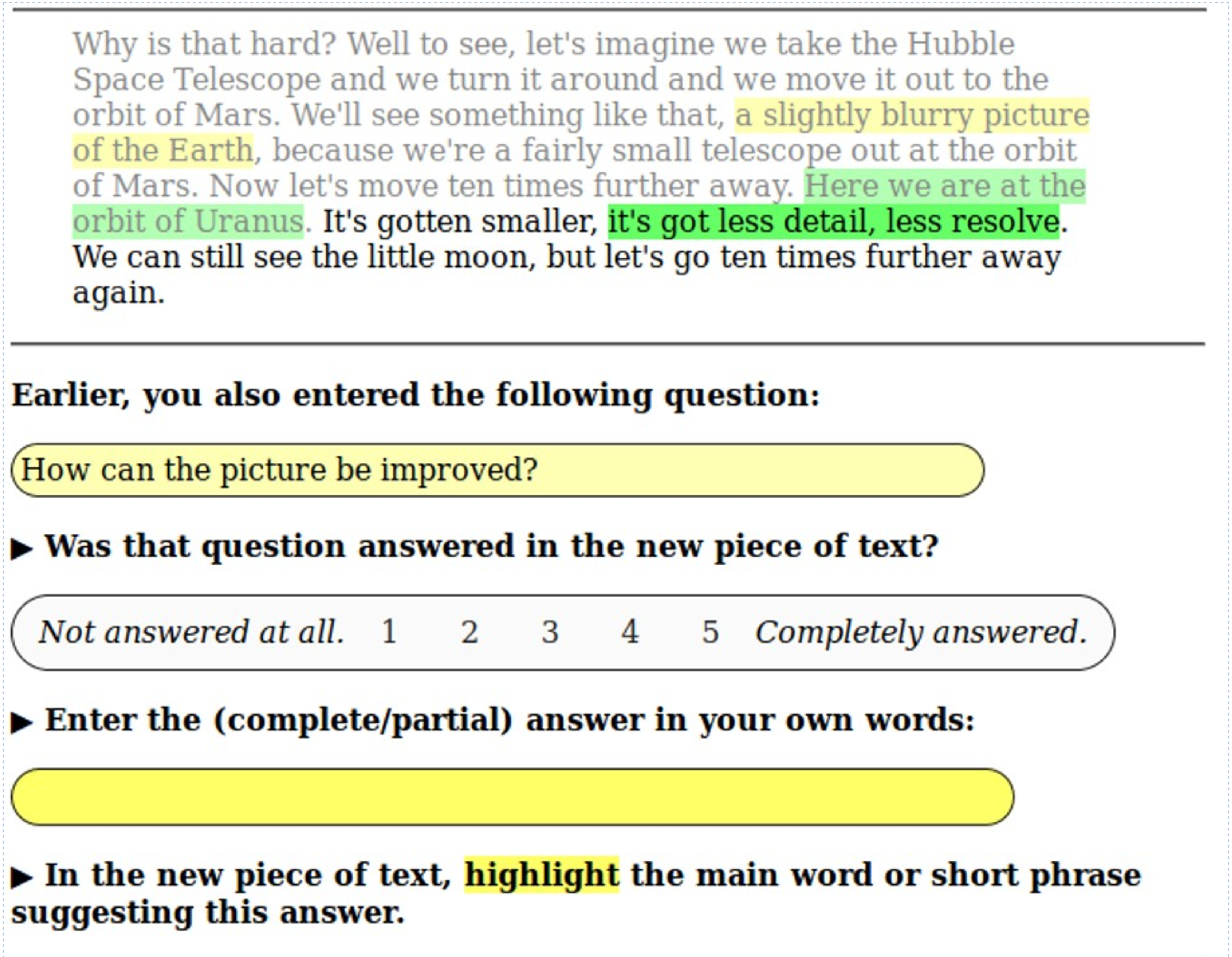
Everybody keeps going on how wee she is

▶ **Please enter a question the text evokes for you at this point.**
(The text so far must *not* yet contain an answer to the question!)

Is she really small?

▶ **In the text, highlight the main word or short phrase that evokes this question.**

Elicitation tool



Why is that hard? Well to see, let's imagine we take the Hubble Space Telescope and we turn it around and we move it out to the orbit of Mars. We'll see something like that, a slightly blurry picture of the Earth, because we're a fairly small telescope out at the orbit of Mars. Now let's move ten times further away. Here we are at the orbit of Uranus. It's gotten smaller, it's got less detail, less resolve. We can still see the little moon, but let's go ten times further away again.

Earlier, you also entered the following question:

How can the picture be improved?

▶ **Was that question answered in the new piece of text?**

Not answered at all. 1 2 3 4 5 *Completely answered.*

▶ **Enter the (complete/partial) answer in your own words:**

▶ **In the new piece of text, highlight the main word or short phrase suggesting this answer.**

Source texts

Source texts

- 6 TED-talks (6975 words) from TED-MDB. Rehbein et al. 2016

Source texts

- 6 TED-talks (6975 words) from TED-MDB. Rehbein et al. 2016
- 2 dialogues (3807 words) from DISCO-SPICE. Zeyrek et al. 2016

Source texts

- 6 TED-talks (6975 words) from TED-MDB.
  Rehbein et al. 2016
- 2 dialogues (3807 words) from DISCO-SPICE.
  Zeyrek et al. 2016
- (1 short story we constructed.)

Source texts

- 6 TED-talks (6975 words) from TED-MDB.

  Rehbein et al. 2016

- 2 dialogues (3807 words) from DISCO-SPICE.

  Zeyrek et al. 2016

- (1 short story we constructed.)

Both come with PDTB-style discourse annotations…

Source texts

- 6 TED-talks (6975 words) from TED-MDB.
  Rehbein et al. 2016
- 2 dialogues (3807 words) from DISCO-SPICE.
  Zeyrek et al. 2016
- (1 short story we constructed.)

Both come with PDTB-style discourse annotations…

Source texts

- 6 TED-talks (6975 words) from TED-MDB.
  Rehbein et al. 2016
- 2 dialogues (3807 words) from DISCO-SPICE.
  Zeyrek et al. 2016
- (1 short story we constructed.)

Both come with PDTB-style discourse annotations...

Could evoked questions provide a useful empirical window on discourse structure?

Source texts

- 6 TED-talks (6975 words) from TED-MDB.

Rehbein et al. 2016

- 2 dialogues (3807 words) from DISCO-SPICE.

Zeyrek et al. 2016

- (1 short story we constructed.)

Both come with PDTB-style discourse annotations…

Is discourse structure more explicitly signaled in places where it is less predictable?

Data gathering

Data gathering

- We cut the texts into overlapping excerpts of up to 18 sentences.

Data gathering

- We cut the texts into overlapping excerpts of up to 18 sentences.
- Probe for evoked questions/answers every 2 sentences

Data gathering

- We cut the texts into overlapping excerpts of up to 18 sentences.
- Probe for evoked questions/answers every 2 sentences
  - Ask up to two times whether evoked question is answered.

Data gathering

- We cut the texts into overlapping excerpts of up to 18 sentences.
- Probe for evoked questions/answers every 2 sentences
  - Ask up to two times whether evoked question is answered.
- 111 workers from Mechanical Turk, 6 excerpts each, at least 5 workers per probe.

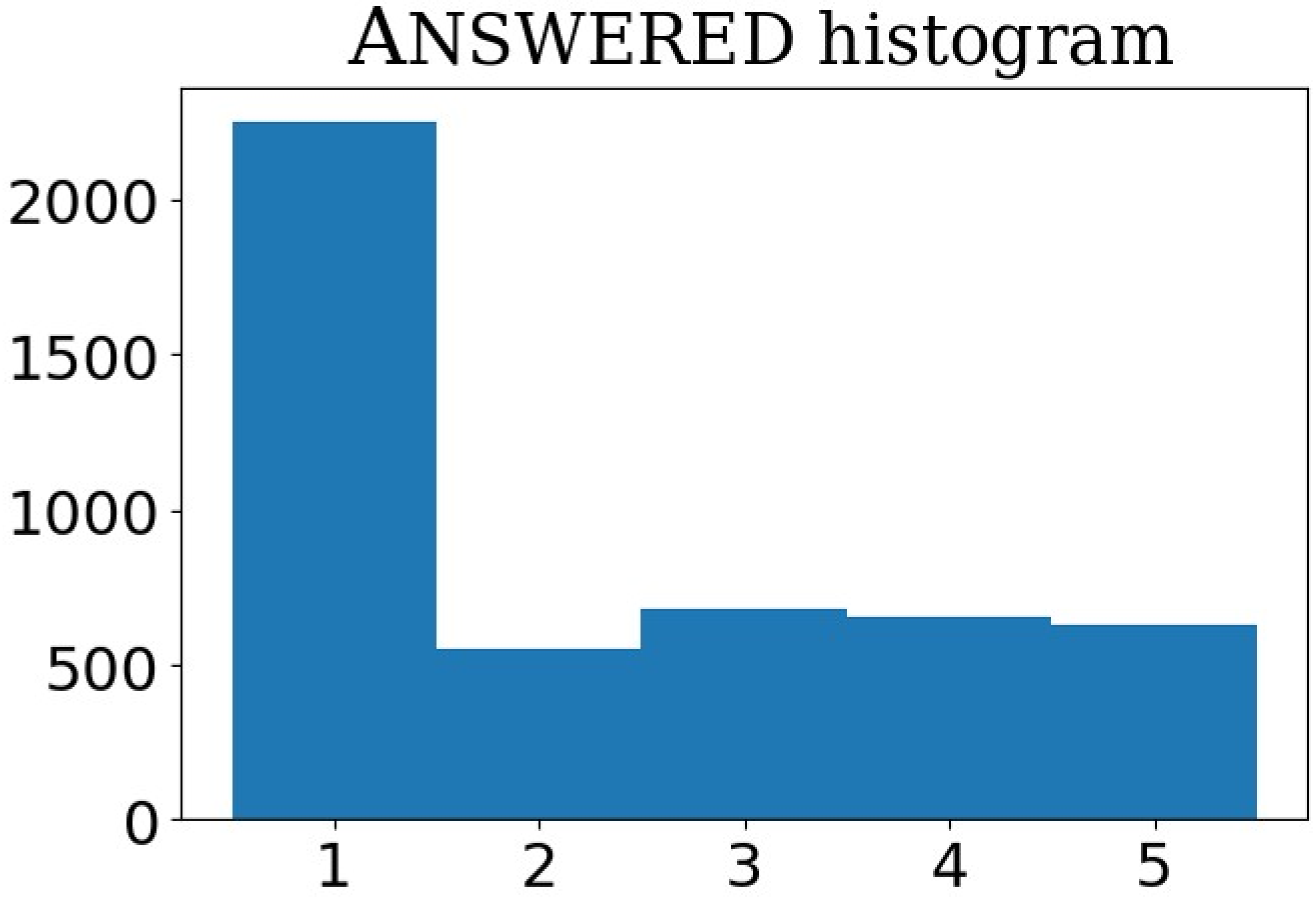
Data gathering

- We cut the texts into overlapping excerpts of up to 18 sentences.
- Probe for evoked questions/answers every 2 sentences
  - Ask up to two times whether evoked question is answered.
- 111 workers from Mechanical Turk, 6 excerpts each, at least 5 workers per probe.
- 50 mins (estim.); reward $8.50.

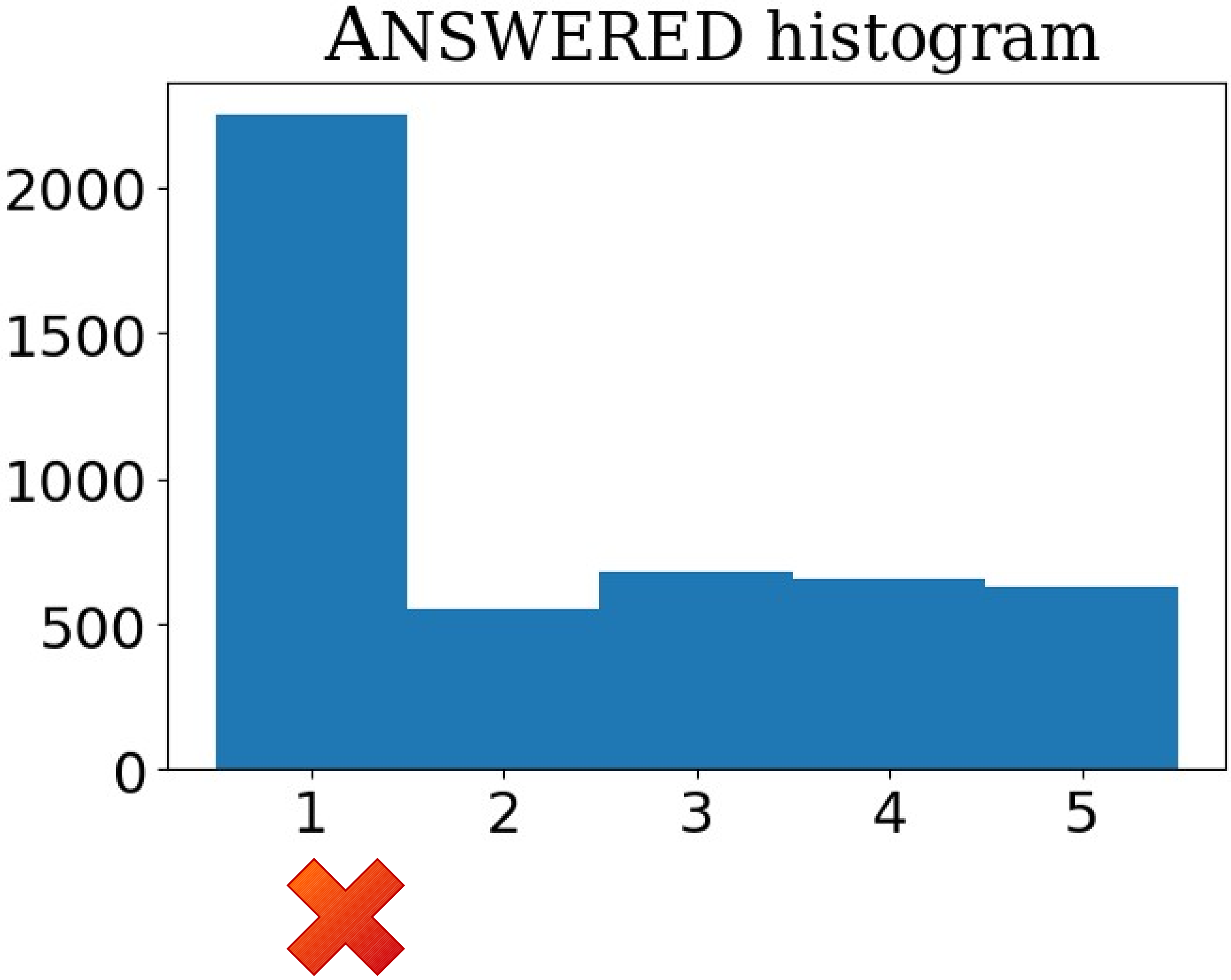
Data gathering

- We cut the texts into overlapping excerpts of up to 18 sentences.
- Probe for evoked questions/answers every 2 sentences
  - Ask up to two times whether evoked question is answered.
- 111 workers from Mechanical Turk, 6 excerpts each, at least 5 workers per probe.
- 50 mins (estim.); reward $8.50.
- **Result:** 863 probes; 4765 questions; 1965 answers.

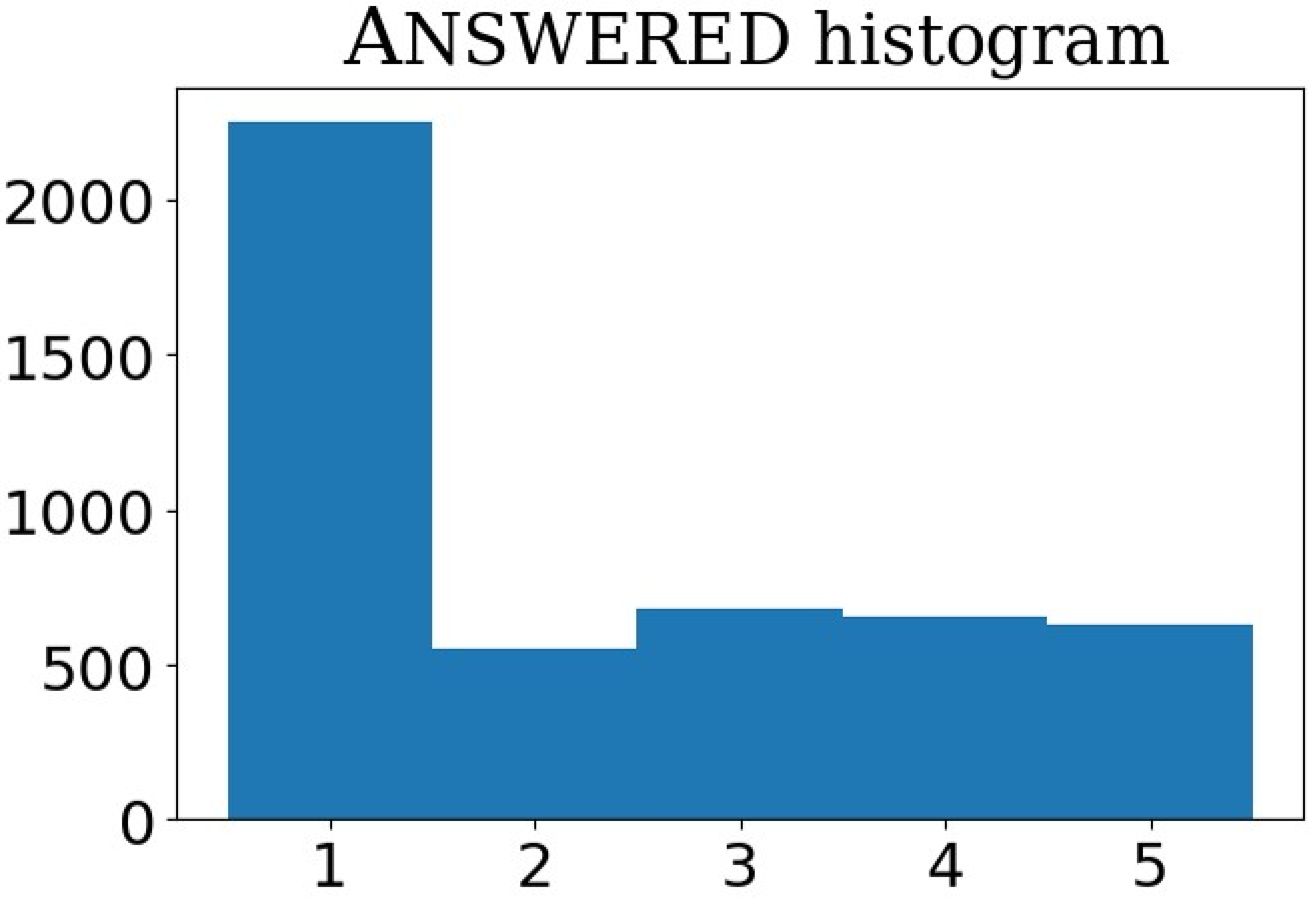
How many questions get answered?



How many questions get answered?

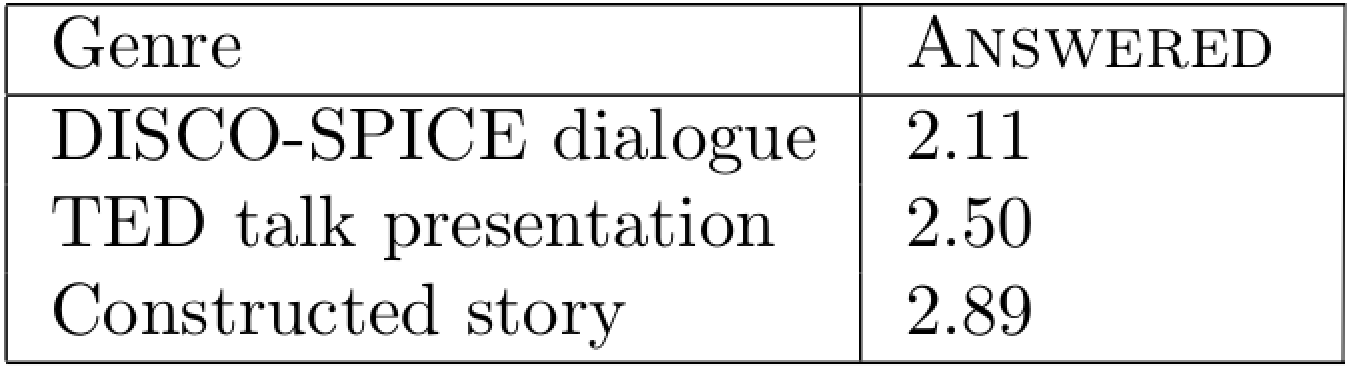


How many questions get answered?



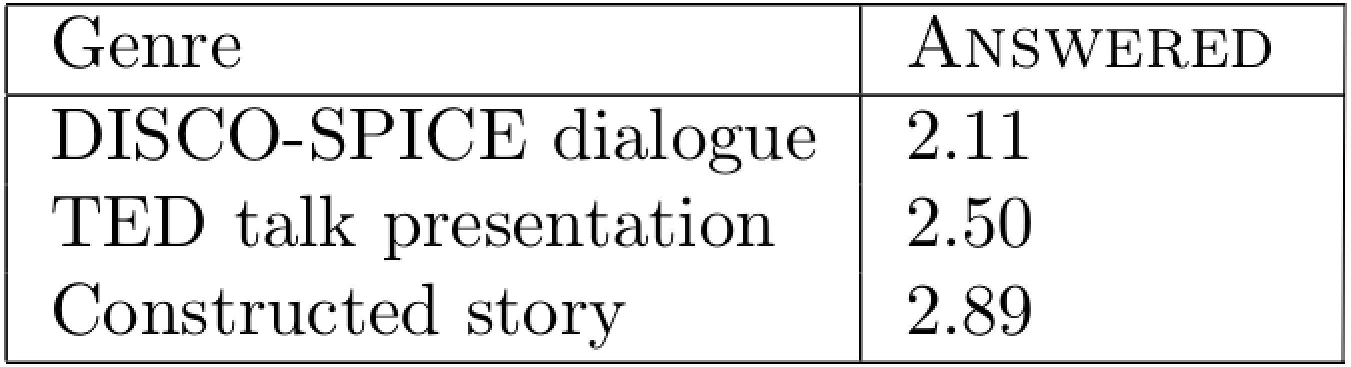
✗ ✓

Mean ANSWERED per genre



Genre	ANSWERED
DISCO-SPICE dialogue	2.11
TED talk presentation	2.50
Constructed story	2.89

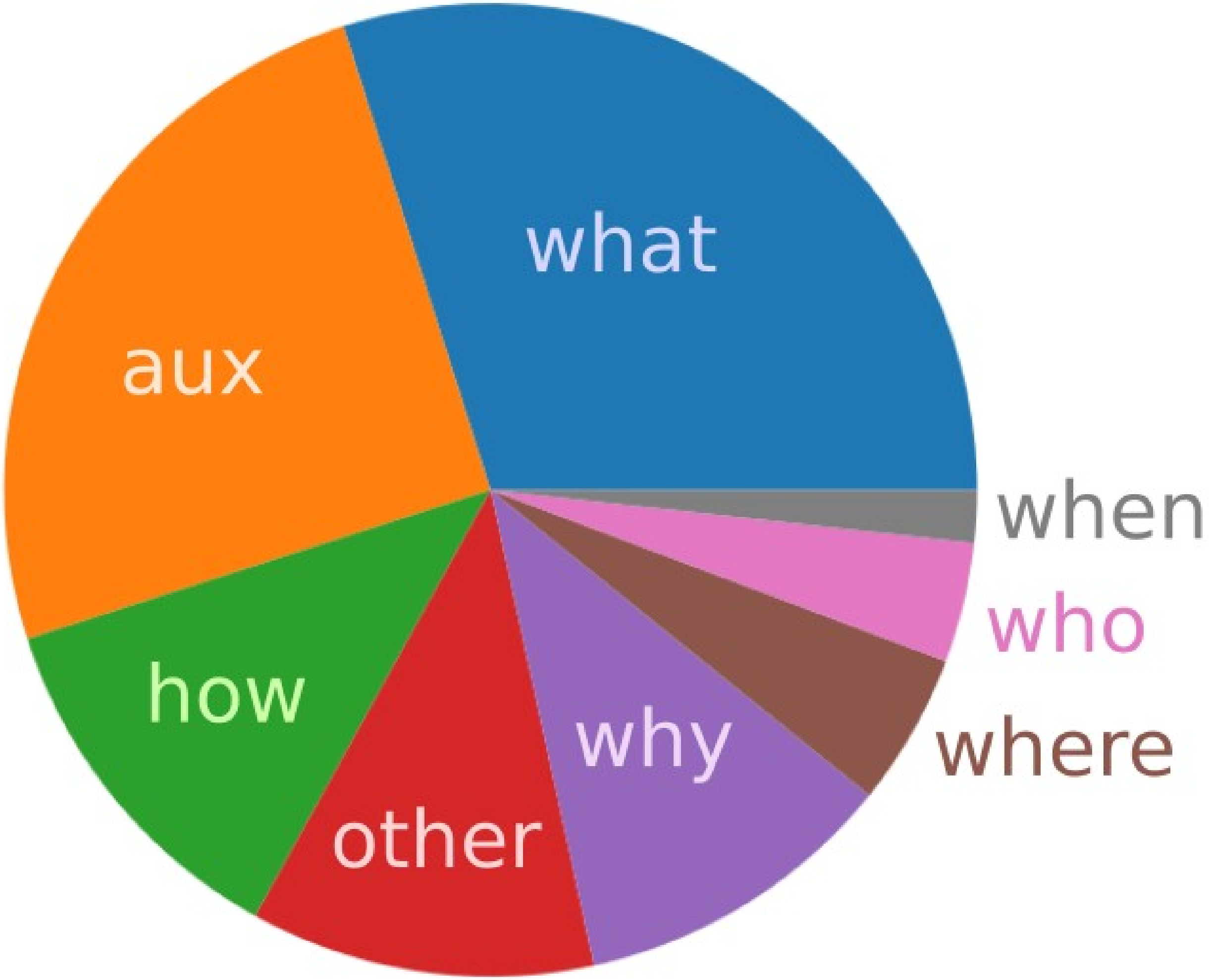
Mean ANSWERED per genre



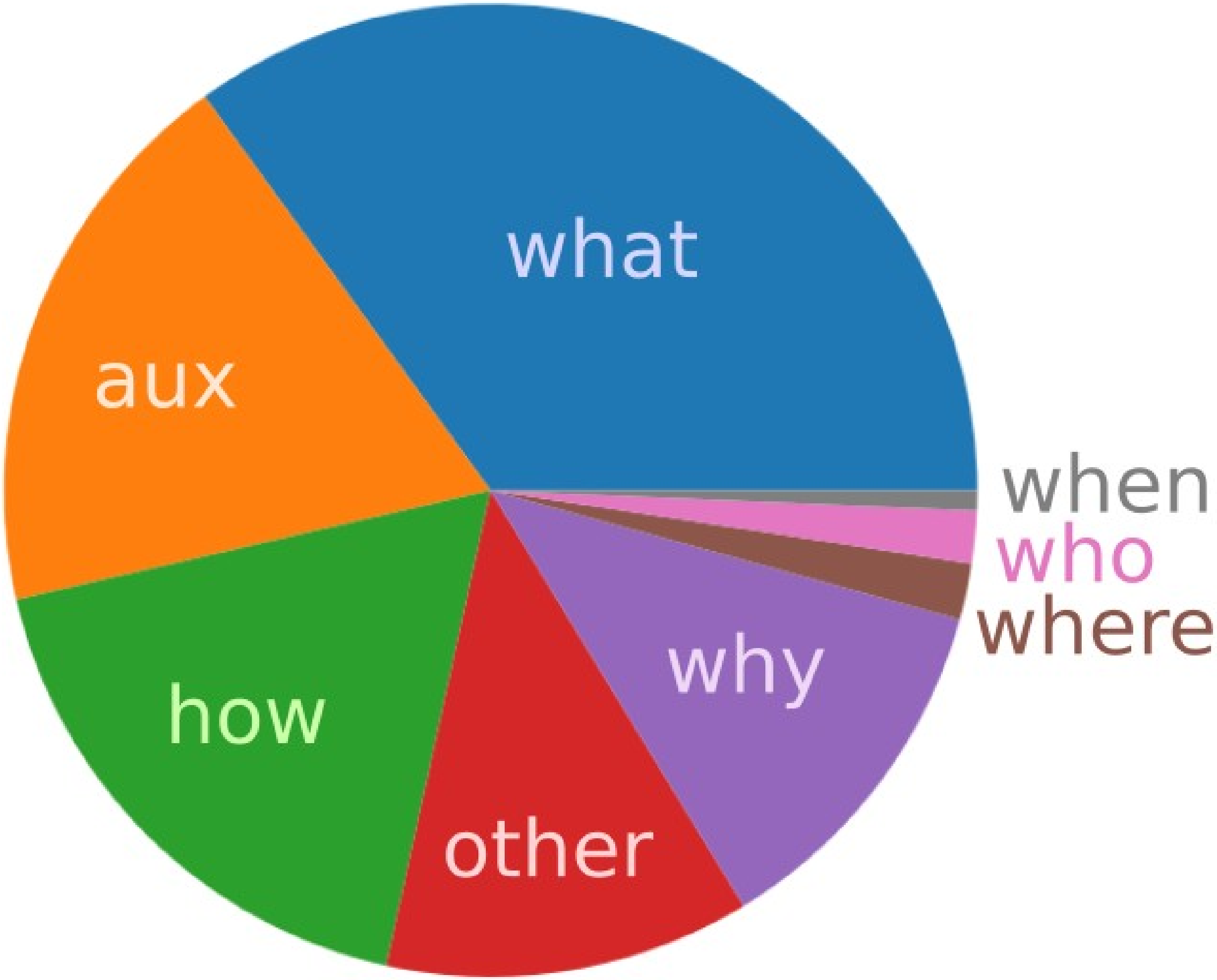
Genre	ANSWERED
DISCO-SPICE dialogue	2.11
TED talk presentation	2.50
Constructed story	2.89

- ANSWERED score of evoked question = ANSWERED score of its best answer (scale 1-5).

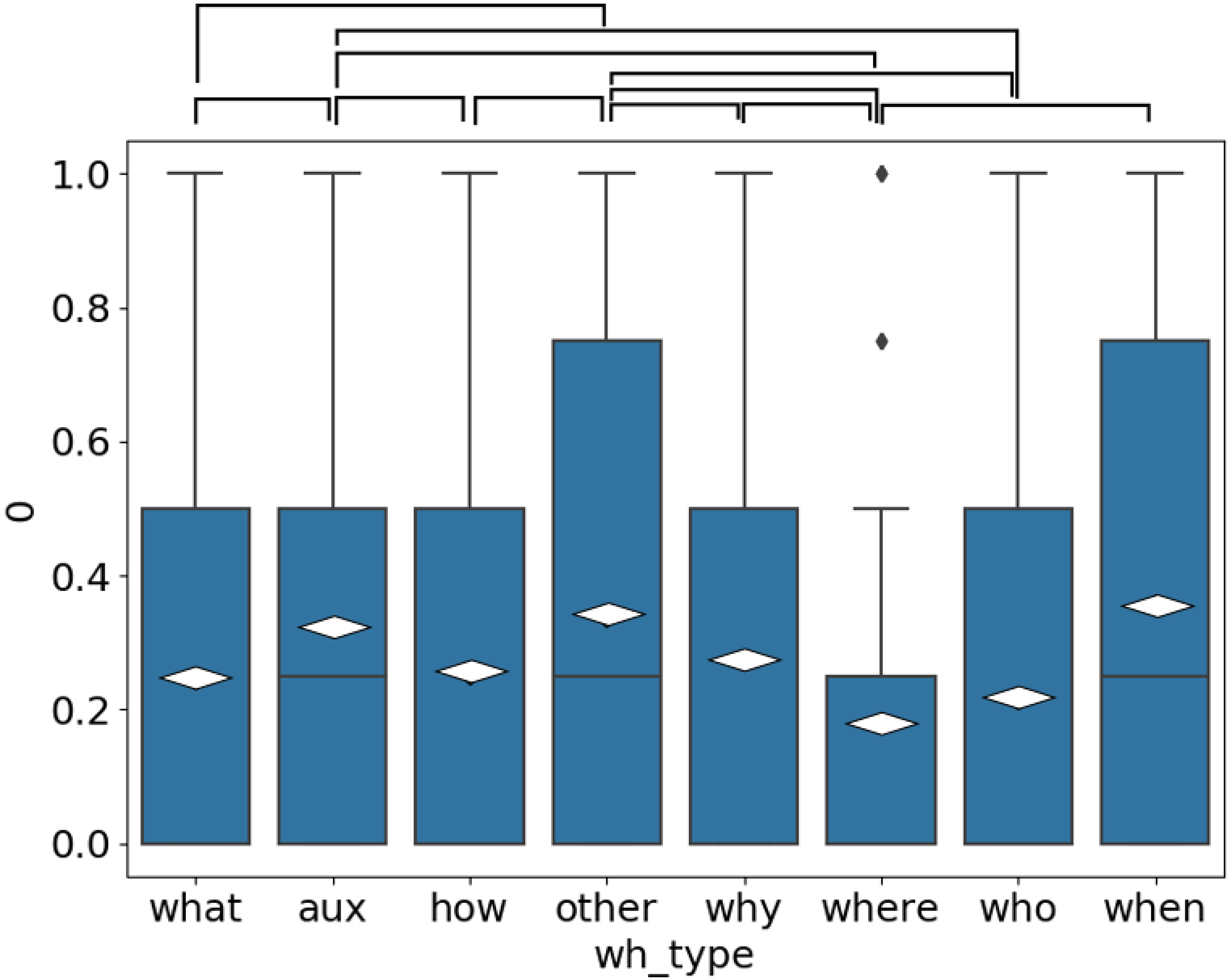
Question types: DISCO-SPICE



Question types: TED-talks



ANSWERED per Q-type: DISCO-SPICE



Research questions

Could evoked questions provide a useful empirical window on discourse structure?

- In particular, can we use them to answer:

Is discourse structure more explicitly signaled in places where it is less predictable?

Research questions

Could evoked questions provide a useful empirical window on discourse structure?

Yes, perhaps they could.

- In particular, can we use them to answer:

Is discourse structure more explicitly signaled in places where it is less predictable?

Research questions

Could evoked questions provide a useful empirical window on discourse structure?

Yes, perhaps they could.

- In particular, can we use them to answer:

Is discourse structure more explicitly signaled in places where it is less predictable?

No idea?

Zooming in on the TED-talks

Zooming in on the TED-talks

- TED-MDB's existing discourse annotations have better coverage.

Zooming in on the TED-talks

- TED-MDB's existing discourse annotations have better coverage.

DISCO-SPICE covers only within-utterance relations :(

Zooming in on the TED-talks

- TED-MDB's existing discourse annotations have better coverage.

  DISCO-SPICE covers only within-utterance relations :(

- The TED-talks seem more suitable for our (current) elicitation task anyway.

Zooming in on the TED-talks

- TED-MDB's existing discourse annotations have better coverage.

  DISCO-SPICE covers only within-utterance relations :(

- The TED-talks seem more suitable for our (current) elicitation task anyway.

- So we further annotated and analyzed our TED-portion, and submitted it to LREC as **TED-Q**.

TED-Q + TED-MDB = ♥

TED-Q + TED-MDB = ♥

Is discourse structure more explicitly signaled in places where it is less predictable?

TED-Q + TED-MDB = ♥

Is discourse structure more explicitly signaled in places where it is less predictable?

- TED-Q gives us a measure of discourse structure predictability (ANSWERED).

TED-Q + TED-MDB = ♥

Is discourse structure more explicitly signaled in places where it is less predictable?

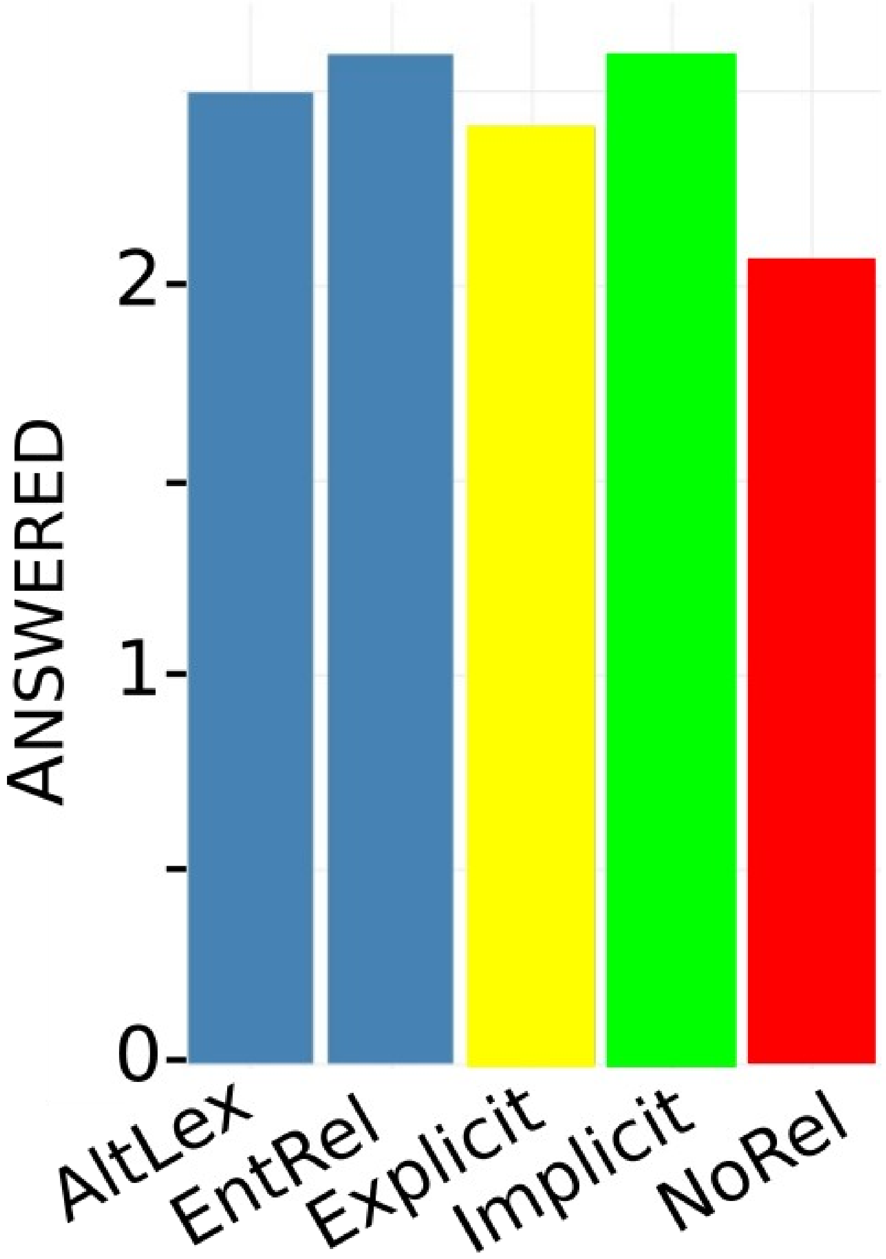
- TED-Q gives us a measure of discourse structure predictability (ANSWERED).
- TED-MDB tells us what the discourse structure actually is…

TED-Q + TED-MDB = ♥

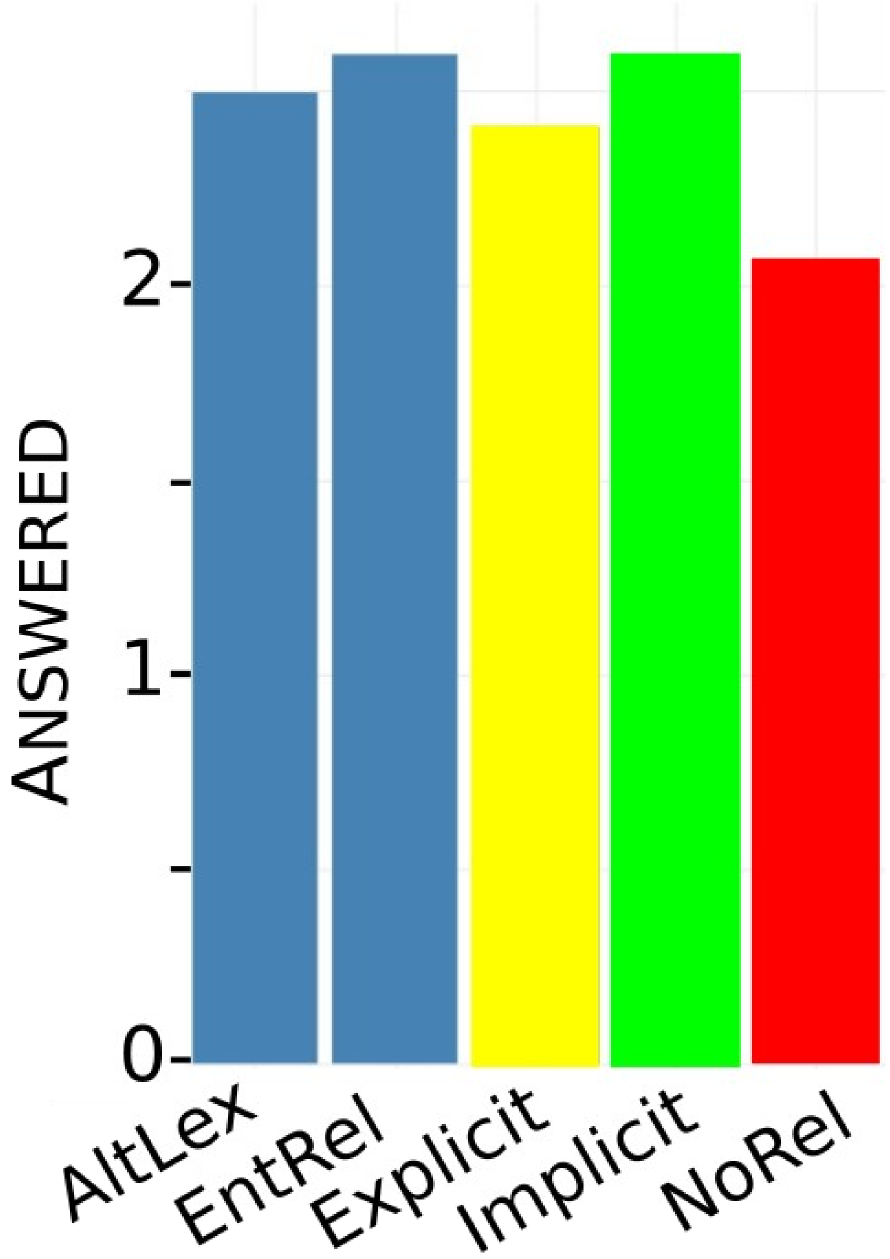
Is discourse structure more explicitly signaled in places where it is less predictable?

- TED-Q gives us a measure of discourse structure predictability (ANSWERED).
- TED-MDB tells us what the discourse structure actually is…
- … and whether it is made explicit or not (connectives).

Explicit vs. implicit Discourse Rel.



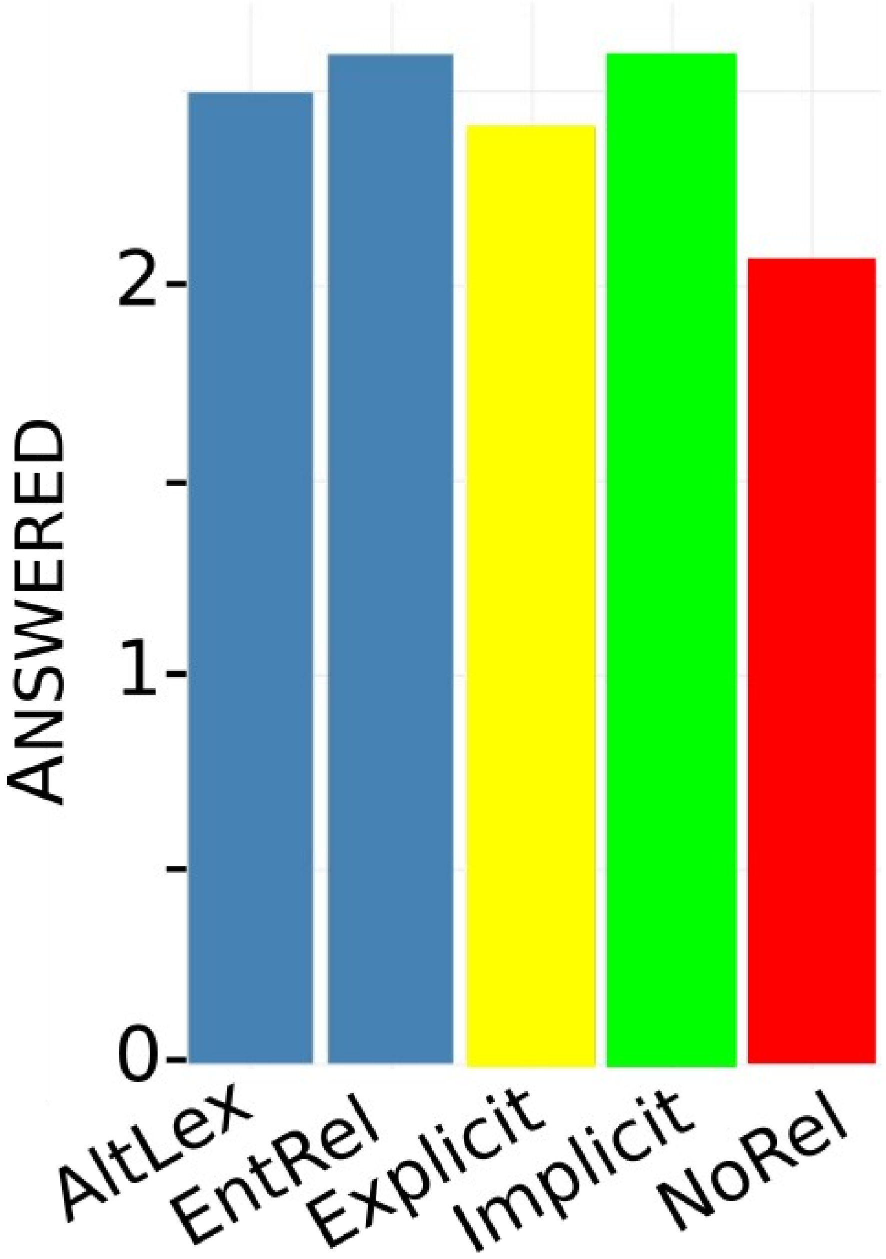
Explicit vs. implicit Discourse Rel.



ANSWERED rating:

- Explicit < Implicit (t(1580)=2.39, p=.016)

Explicit vs. implicit Discourse Rel.

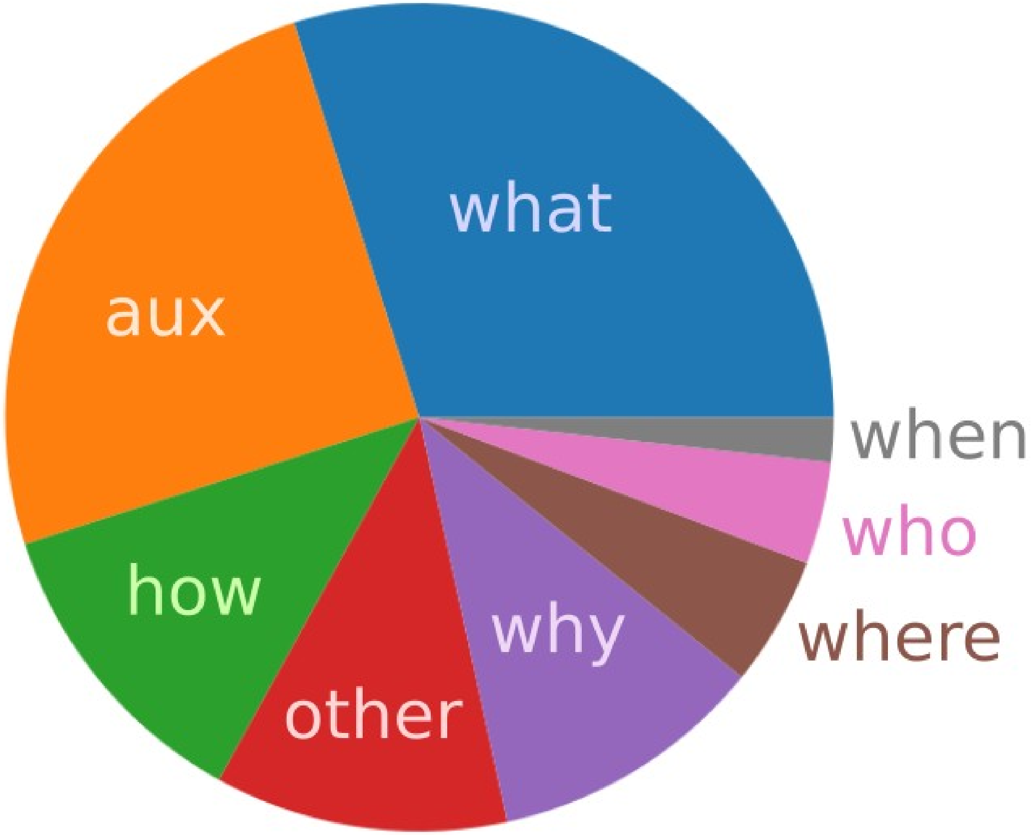


ANSWERED rating:

- Explicit < Implicit (t(1580)=2.39, p=.016)
- NoRel < others (t(2219)=4.71, p<.0001)

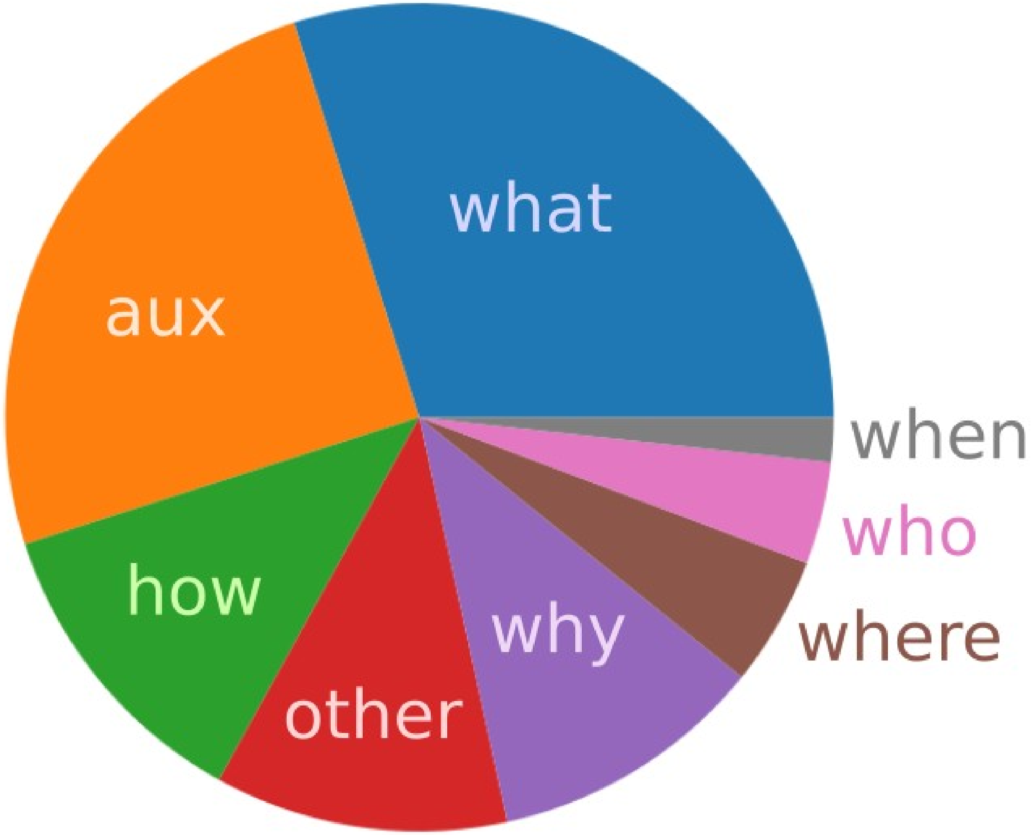
Question types ~ relations?

Question types ~ relations?



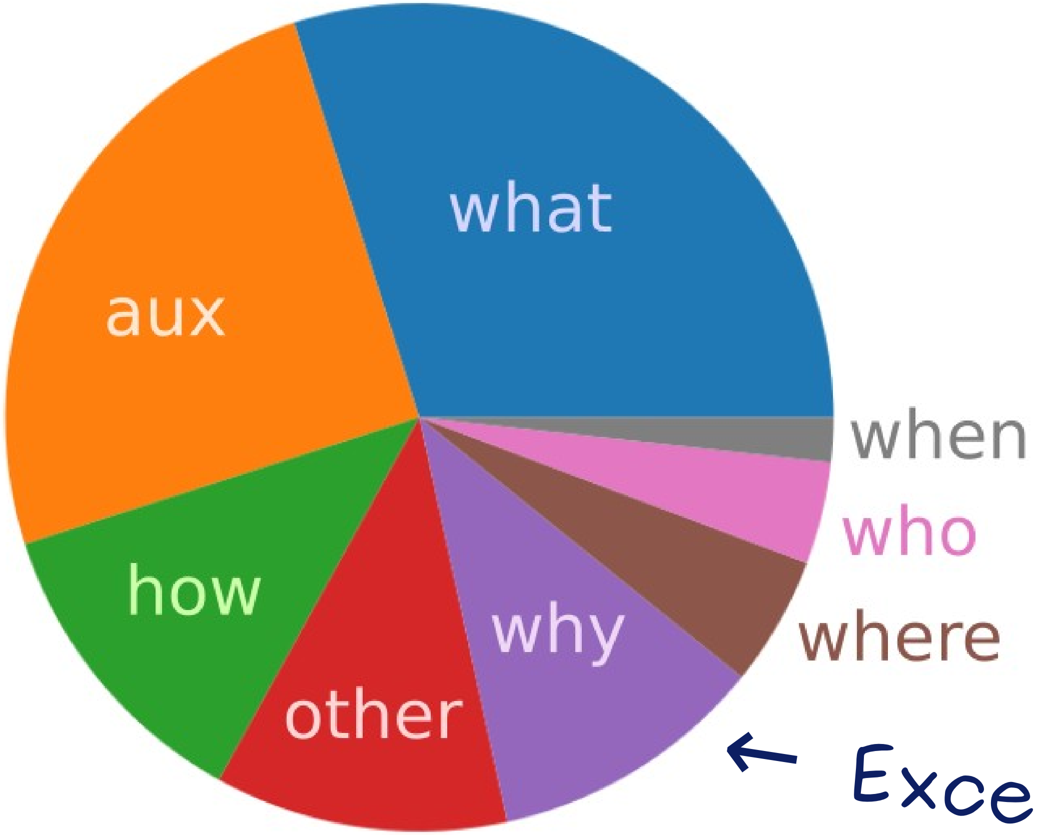
Question types ~ relations?

No significant effects…



Question types ~ relations?

No significant effects…



← Except here!
12%→19% (p<.01)

Conclusion

Is discourse structure more explicitly signaled in places where it is less predictable?

Conclusion

Is discourse structure more explicitly signaled in places where it is less predictable?

Yes it is!

Conclusion

Is discourse structure more explicitly signaled in places where it is less predictable?

Yes it is!

- Previously confirmed using coarse generalizations, e.g.:

Conclusion

Is discourse structure more explicitly signaled in places where it is less predictable?

Yes it is!

- Previously confirmed using coarse generalizations, e.g.:
  - "Causal relations more predictable."
  - "Continuous relations more predictable."

Conclusion

Is discourse structure more explicitly signaled in places where it is less predictable?

Yes it is!

- Previously confirmed using coarse generalizations, e.g.:
  - "Causal relations more predictable."
  - "Continuous relations more predictable."
- TED-MDB + TED-Q enables a new kind of confirmation:

Conclusion

Is discourse structure more explicitly signaled in places where it is less predictable?

Yes it is!

- Previously confirmed using coarse generalizations, e.g.:
  - "Causal relations more predictable."
  - "Continuous relations more predictable."
- TED-MDB + TED-Q enables a new kind of confirmation:
  - without prior assumptions about predictability;

Conclusion

Is discourse structure more explicitly signaled in places where it is less predictable?

Yes it is!

- Previously confirmed using coarse generalizations, e.g.:
  - "Causal relations more predictable."
  - "Continuous relations more predictable."
- TED-MDB + TED-Q enables a new kind of confirmation:
  - without prior assumptions about predictability;
  - purely data-driven, quantitative, context-sensitive.

Did people pose the same question?

[…] Many of you might be wondering why anyone would choose a life like this, under the thumb of discriminatory laws

Did people pose the same question?

[…] Many of you might be wondering why anyone would choose a life like this, under the thumb of discriminatory laws

Why would people want to live this way?

Did people pose the same question?

[...] Many of you might be wondering why anyone would choose a life like this, under the thumb of discriminatory laws

Why would people want to live this way?
What are some reasons the homeless might give
for living like they do?

Did people pose the same question?

[…] Many of you might be wondering why anyone would choose a life like this, under the thumb of discriminatory laws

Why would people want to live this way?
What are some reasons the homeless might give for living like they do?
Why would anyone live such a dangerous life?

Did people pose the same question?

[...] Many of you might be wondering why anyone would choose a life like this, under the thumb of discriminatory laws

Why would people want to live this way?
What are some reasons the homeless might give for living like they do?
Why would anyone live such a dangerous life?
Why are these laws like this?

Did people pose the same question?

[...] Many of you might be wondering why anyone would choose a life like this, under the thumb of discriminatory laws

Why would people want to live this way?
What are some reasons the homeless might give for living like they do?
Why would anyone live such a dangerous life?
Why are these laws like this?
Why choose homelessness?

Annotating question relatedness

Annotating question relatedness

- Crowdsourced task:

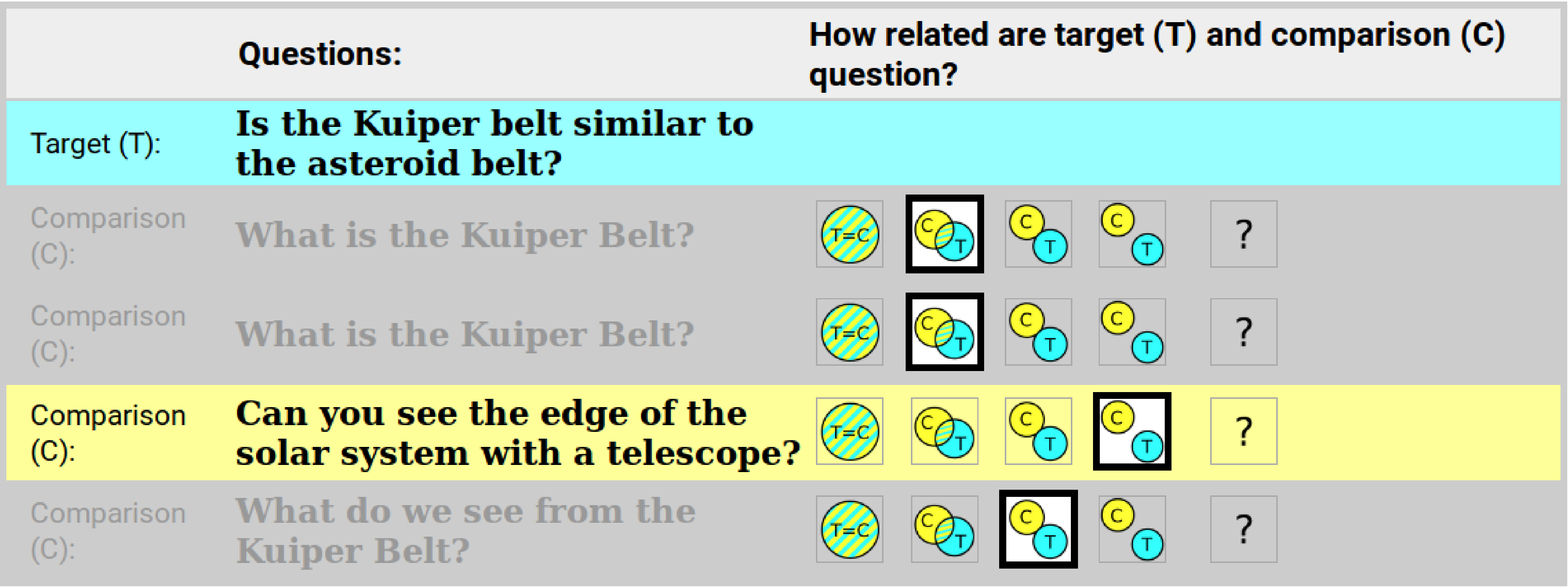
  https://workersandbox.mturk.com/projects/3L2A3M5C0728O1QYNC4O7LNJPOCGQW/tasks

Annotating question relatedness

- **Please read the snippet:**

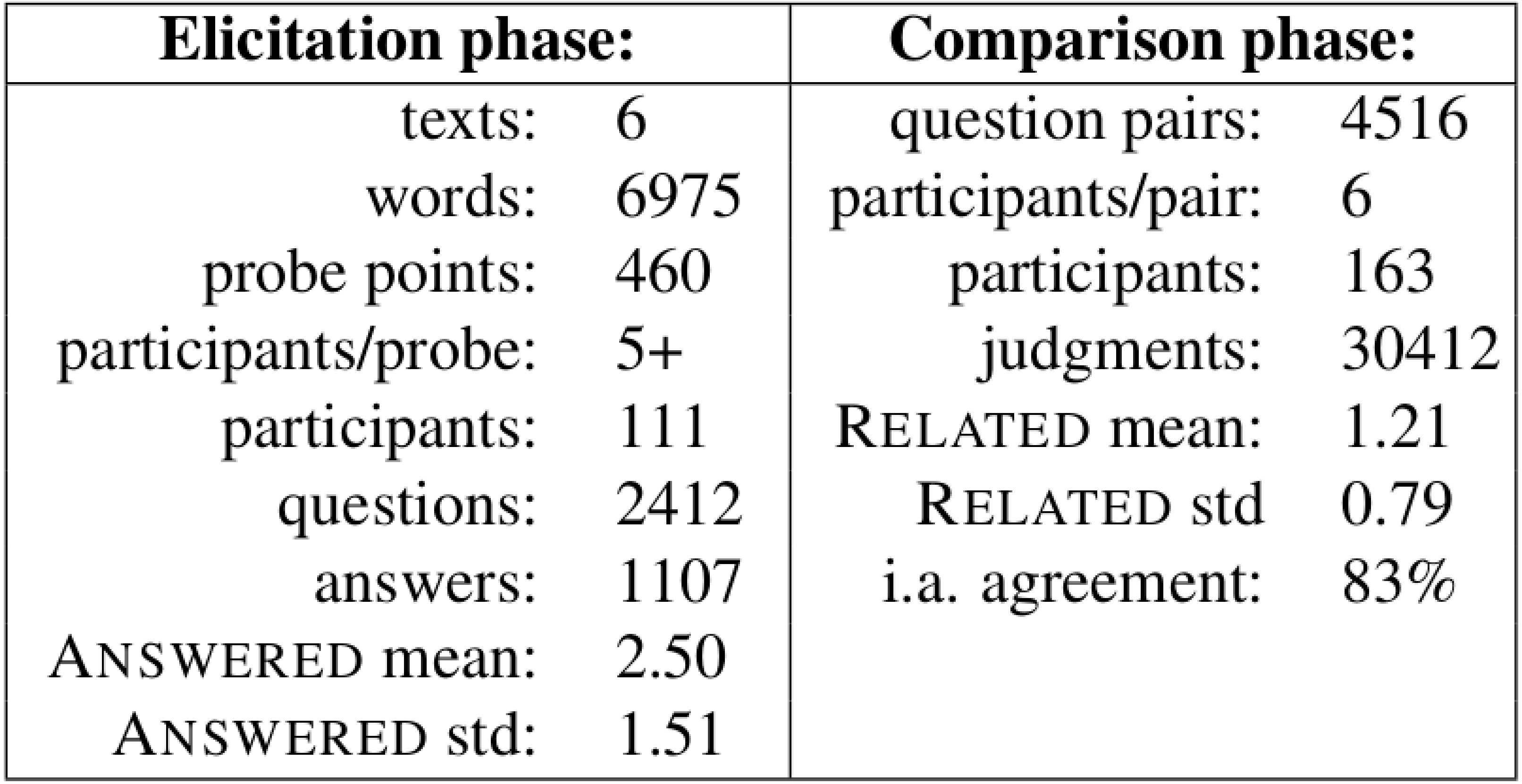
> […] We can still see the little moon, but let's go ten times further away again. Here we are at the edge of the solar system, out at the Kuiper Belt.

- **Next, compare the questions it evoked:**



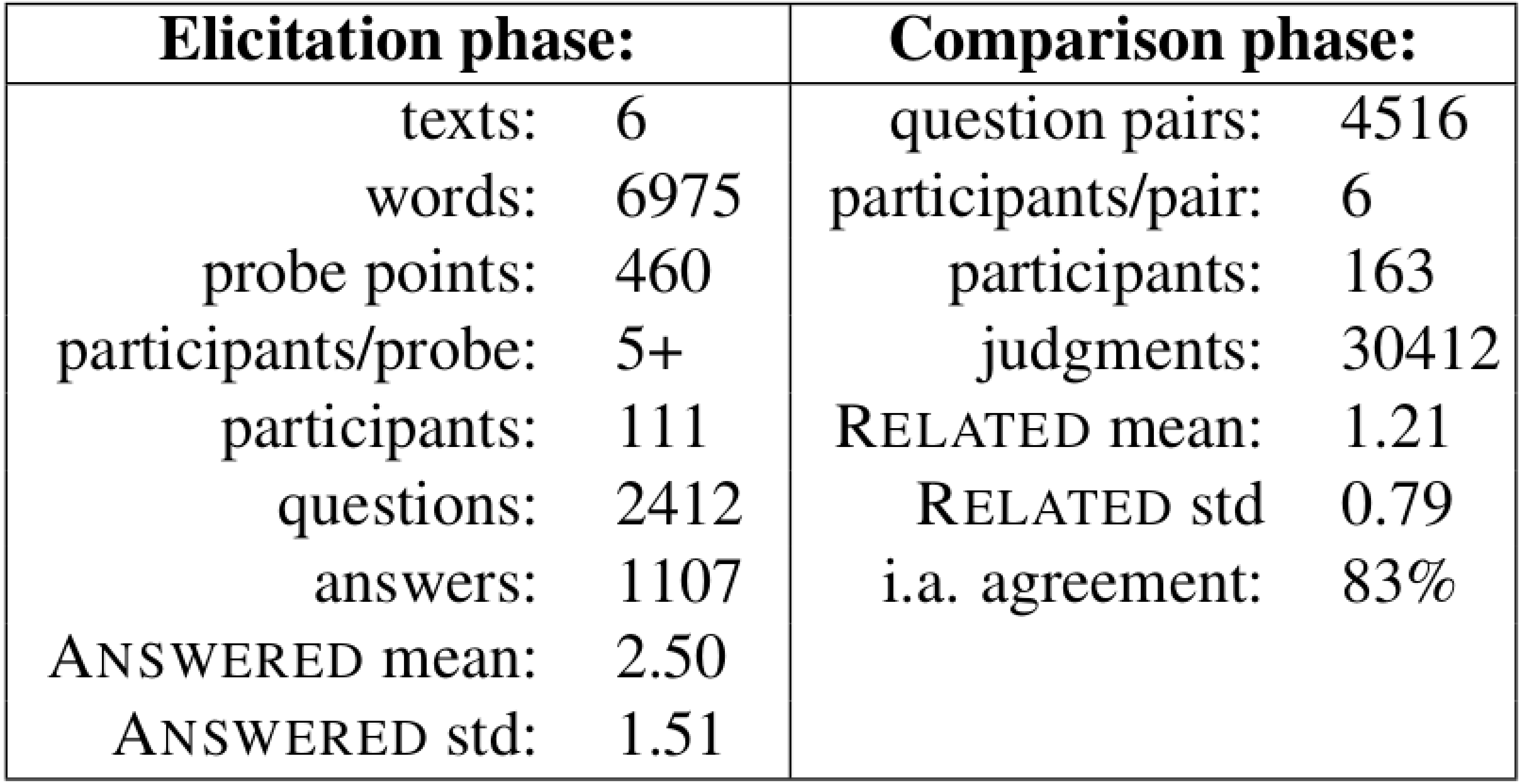
	Questions:	**How related are target (T) and comparison (C) question?**
Target (T):	**Is the Kuiper belt similar to the asteroid belt?**	
Comparison (C):	What is the Kuiper Belt?	T=C [✓ C T] C T C T ?
Comparison (C):	What is the Kuiper Belt?	T=C [✓ C T] C T C T ?
Comparison (C):	**Can you see the edge of the solar system with a telescope?**	T=C C T C T [✓ C T] ?
Comparison (C):	What do we see from the Kuiper Belt?	T=C C T [✓ C T] C T ?

Our resulting dataset: TED-Q



Elicitation phase:		Comparison phase:	
texts:	6	question pairs:	4516
words:	6975	participants/pair:	6
probe points:	460	participants:	163
participants/probe:	5+	judgments:	30412
participants:	111	RELATED mean:	1.21
questions:	2412	RELATED std	0.79
answers:	1107	i.a. agreement:	83%
ANSWERED mean:	2.50		
ANSWERED std:	1.51		

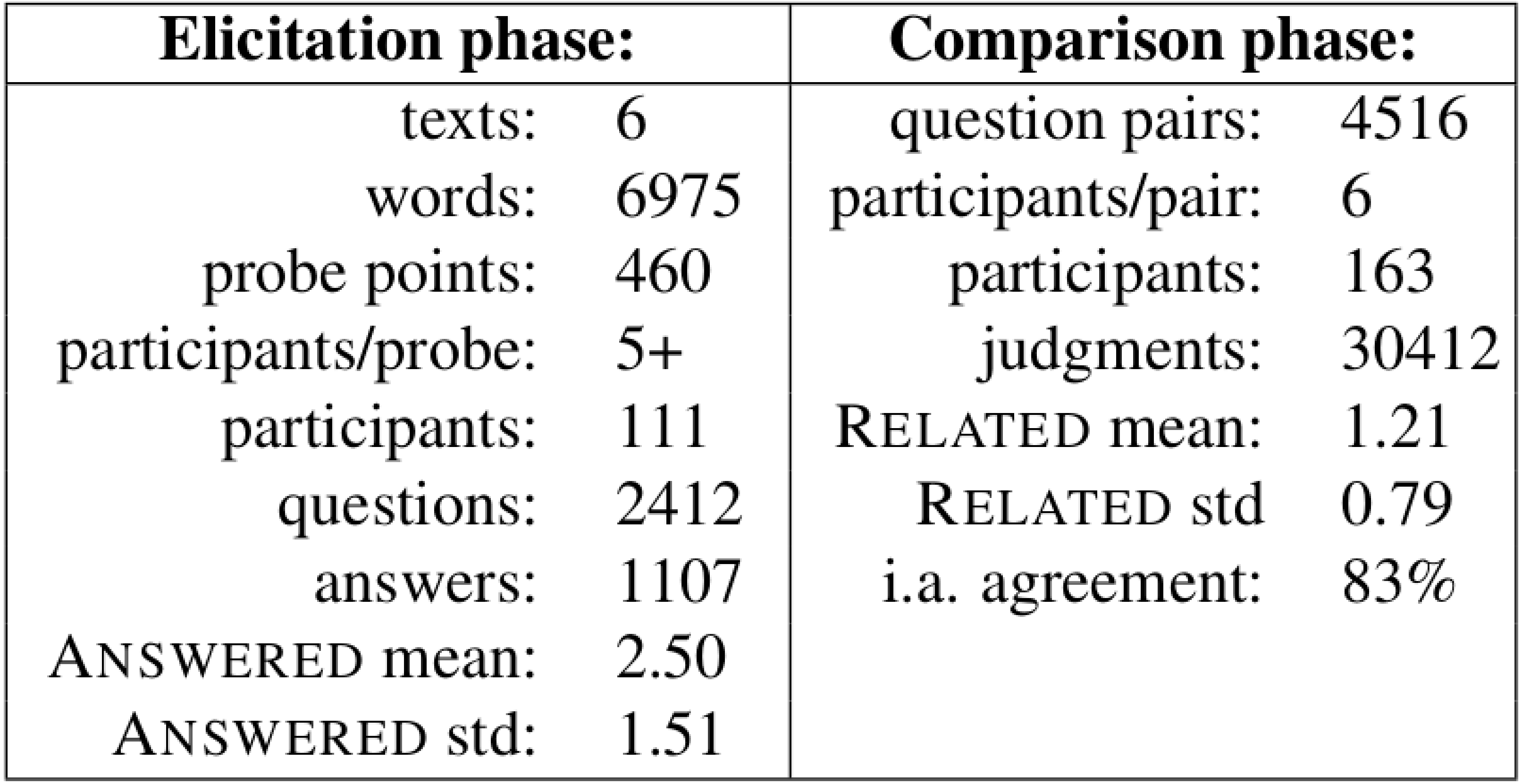
Our resulting dataset: TED-Q



Elicitation phase:		Comparison phase:	
texts:	6	question pairs:	4516
words:	6975	participants/pair:	6
probe points:	460	participants:	163
participants/probe:	5+	judgments:	30412
participants:	111	RELATED mean:	1.21
questions:	2412	RELATED std	0.79
answers:	1107	i.a. agreement:	83%
ANSWERED mean:	2.50		
ANSWERED std:	1.51		

(0…3)

Our resulting dataset: TED-Q

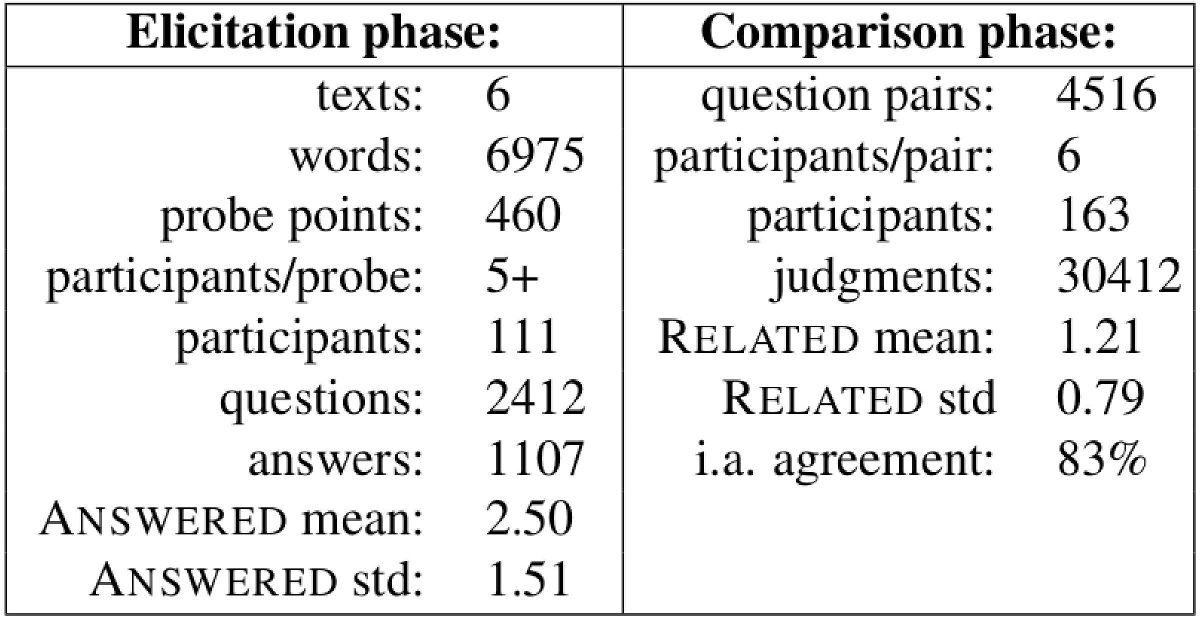


Elicitation phase:		Comparison phase:	
texts:	6	question pairs:	4516
words:	6975	participants/pair:	6
probe points:	460	participants:	163
participants/probe:	5+	judgments:	30412
participants:	111	RELATED mean:	1.21
questions:	2412	RELATED std	0.79
answers:	1107	i.a. agreement:	83%
ANSWERED mean:	2.50		
ANSWERED std:	1.51		

(0…3)

(1…5)

Our resulting dataset: TED-Q

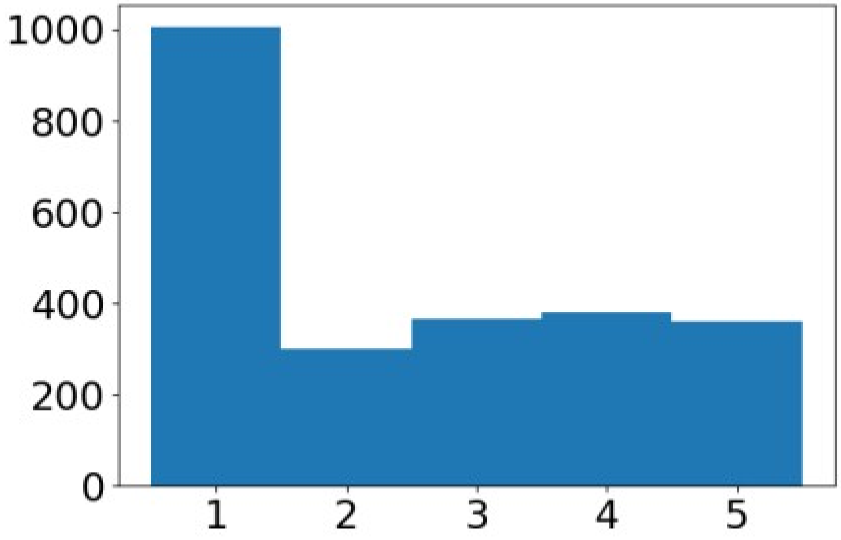


Elicitation phase:		Comparison phase:	
texts:	6	question pairs:	4516
words:	6975	participants/pair:	6
probe points:	460	participants:	163
participants/probe:	5+	judgments:	30412
participants:	111	RELATED mean:	1.21
questions:	2412	RELATED std	0.79
answers:	1107	i.a. agreement:	83%
ANSWERED mean:	2.50		
ANSWERED std:	1.51		

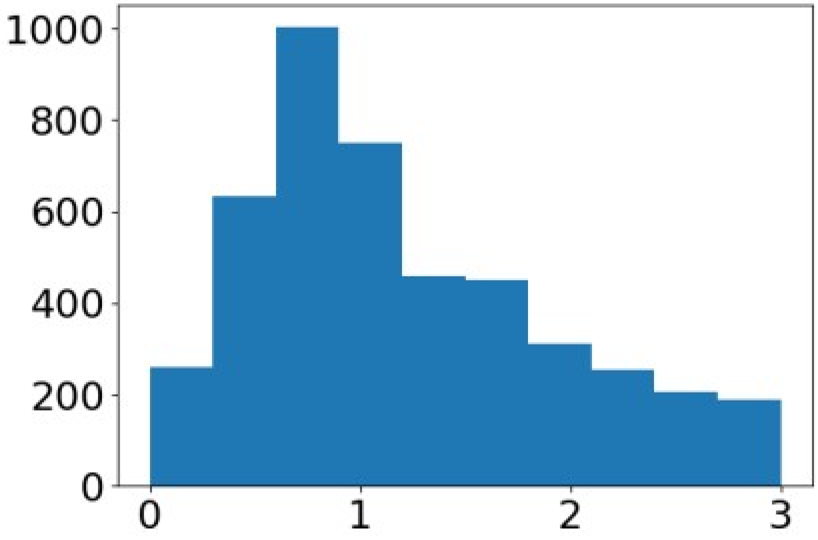
Our resulting dataset: TED-Q

Elicitation phase:		Comparison phase:	
texts:	6	question pairs:	4516
words:	6975	participants/pair:	6
probe points:	460	participants:	163
participants/probe:	5+	judgments:	30412
participants:	111	RELATED mean:	1.21
questions:	2412	RELATED std	0.79
answers:	1107	i.a. agreement:	83%
ANSWERED mean:	2.50		
ANSWERED std:	1.51		

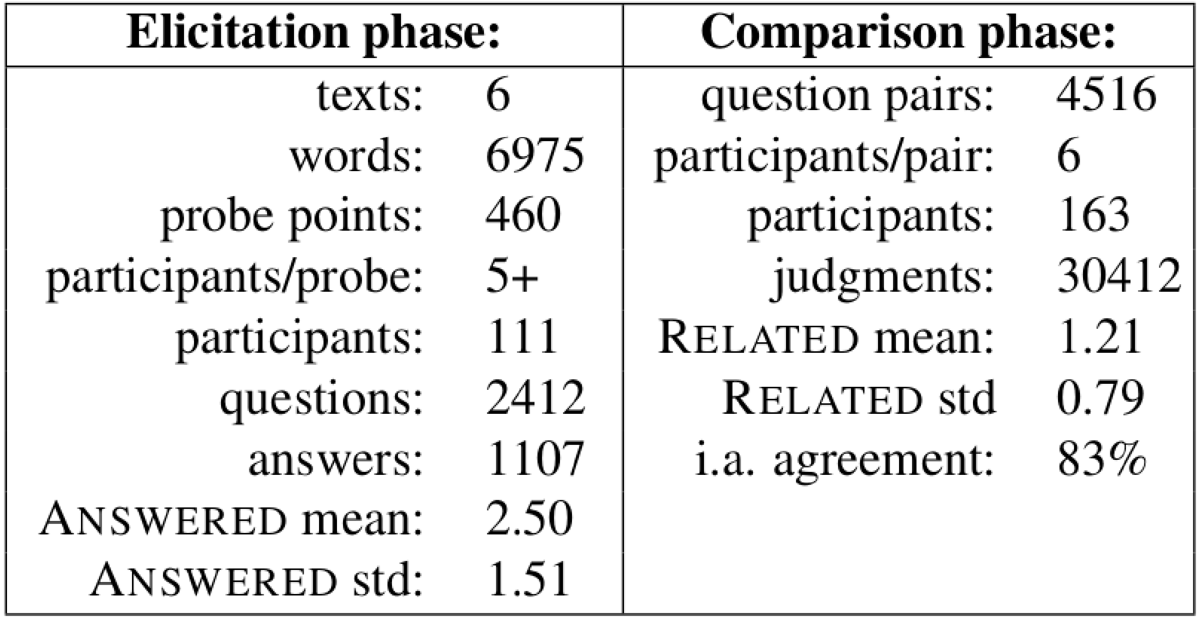
ANSWERED



RELATED

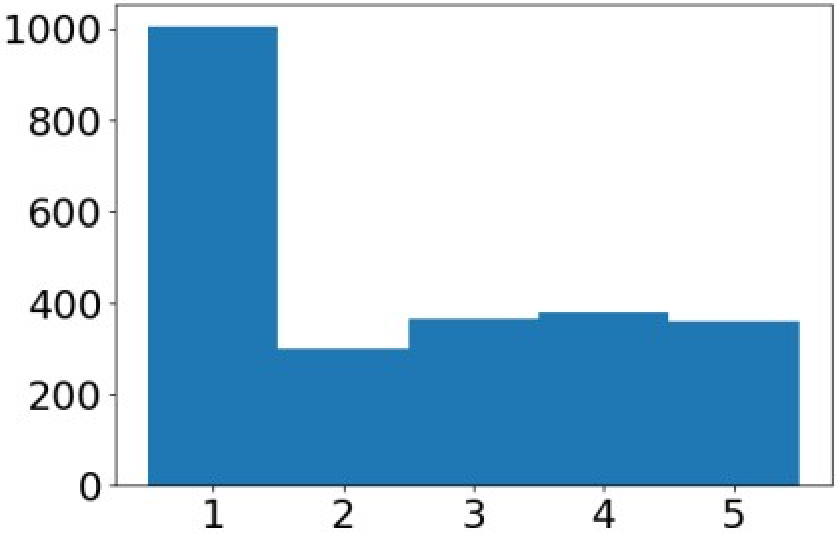


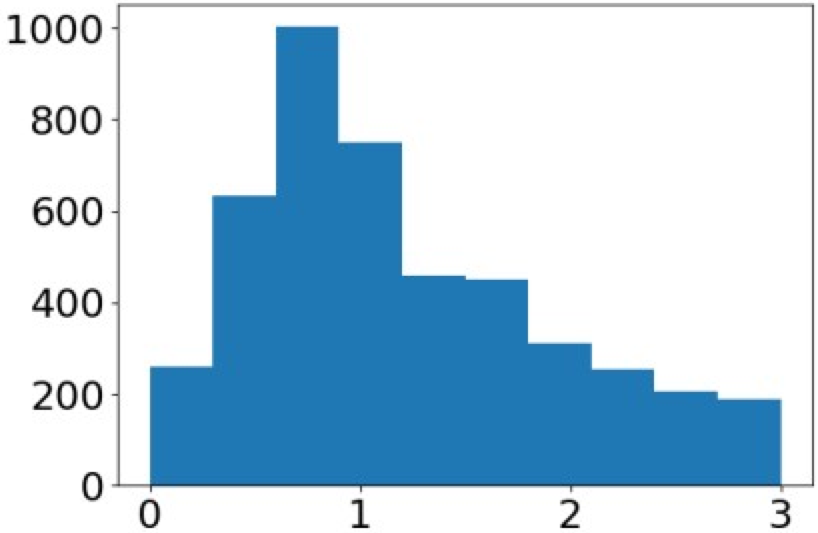
Our resulting dataset: TED-Q



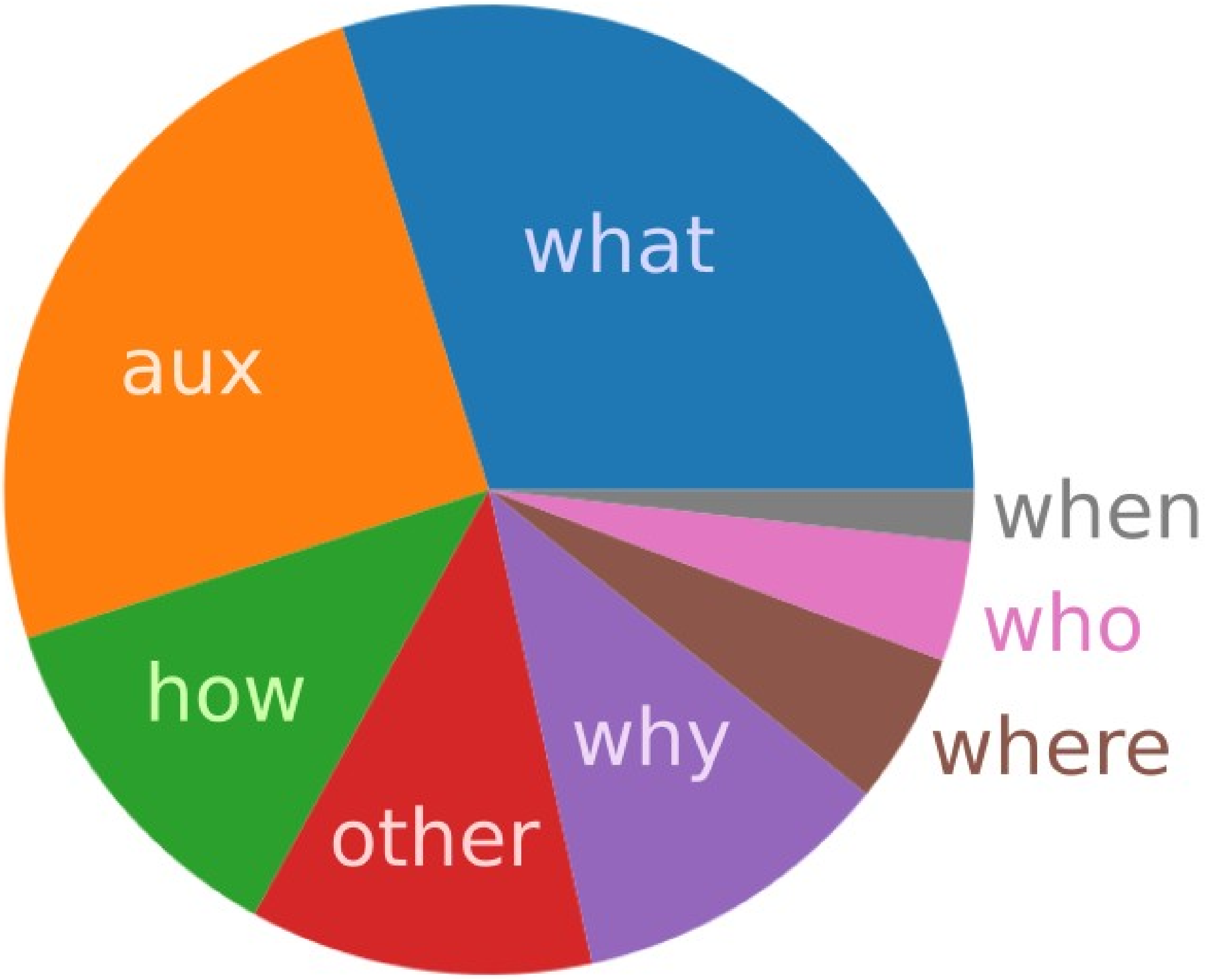
Elicitation phase:		Comparison phase:	
texts:	6	question pairs:	4516
words:	6975	participants/pair:	6
probe points:	460	participants:	163
participants/probe:	5+	judgments:	30412
participants:	111	RELATED mean:	1.21
questions:	2412	RELATED std	0.79
answers:	1107	i.a. agreement:	83%
ANSWERED mean:	2.50		
ANSWERED std:	1.51		

ANSWERED ← Spearman 0.17 → RELATED

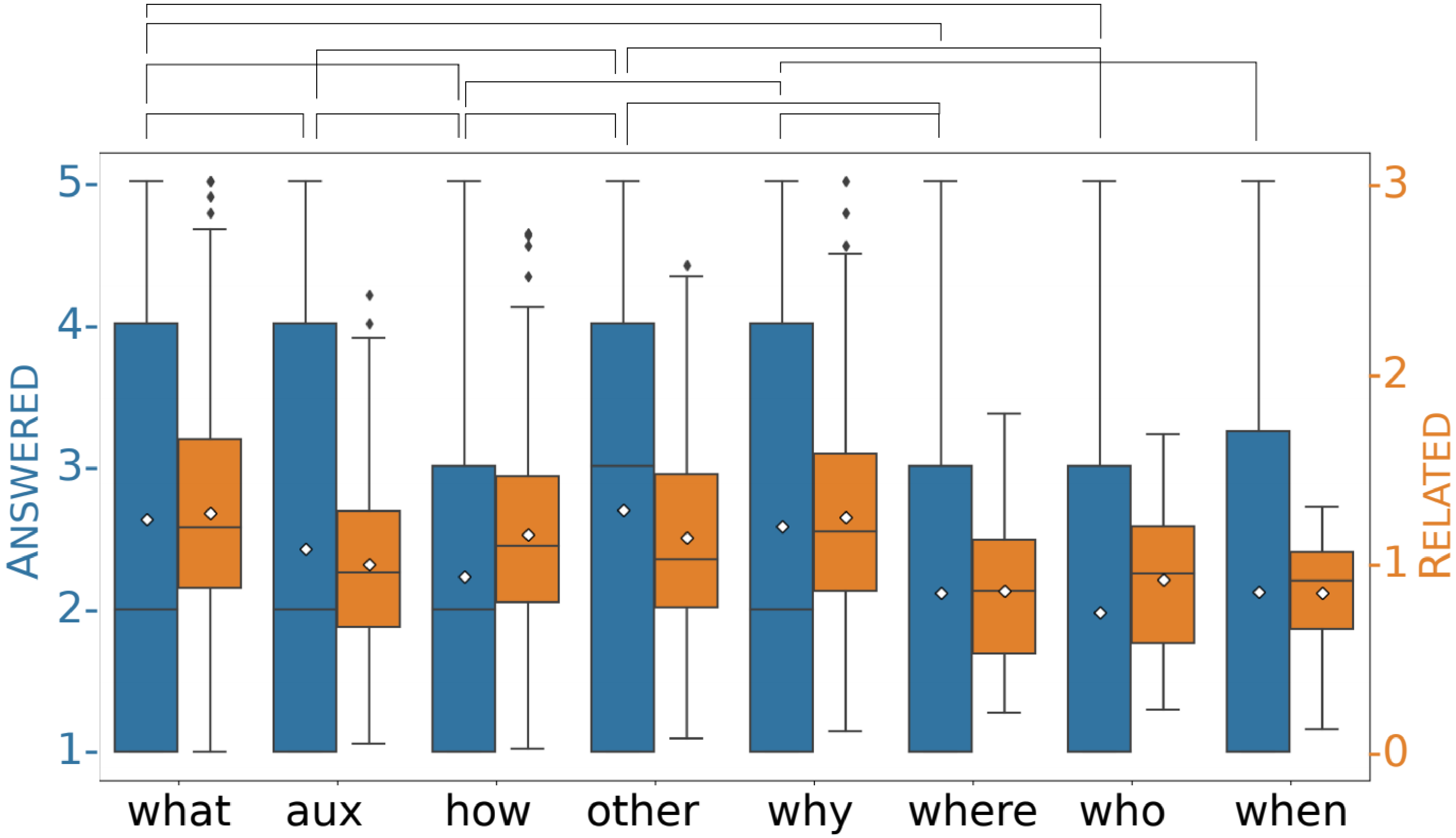




ANSWERED & RELATED per wh-type



ANSWERED & RELATED per wh-type



Funding

This work was supported in part by a Leverhulme Trust Prize in Languages and Literatures to H. Rohde. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154), and from the Spanish State Research Agency (AEI) and the European Regional Development Fund (FEDER, UE) (project PGC2018-094029-A-I00). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

erc
European Research Council
Established by the European Commission