Backtranslation Feedback Improves User Confidence in MT, Not Quality

Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, Lisa Yankovskaya

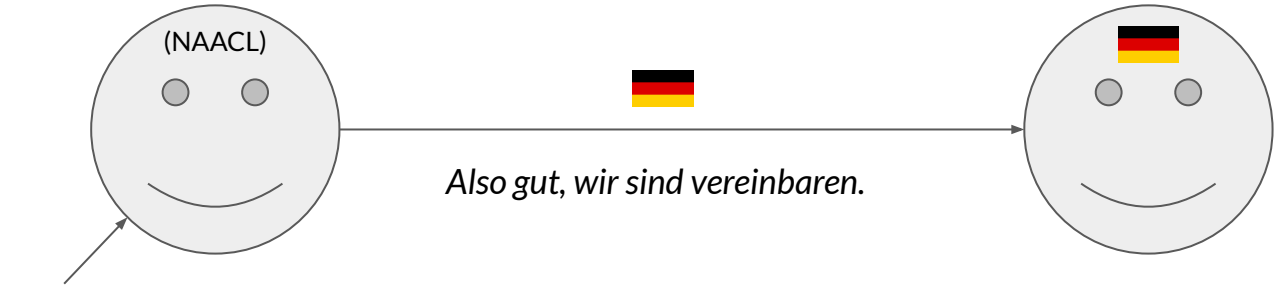
Charles University
University of Edinburgh
University of Sheffield
University of Wolverhampton
University of Tartu

NAACL 2021

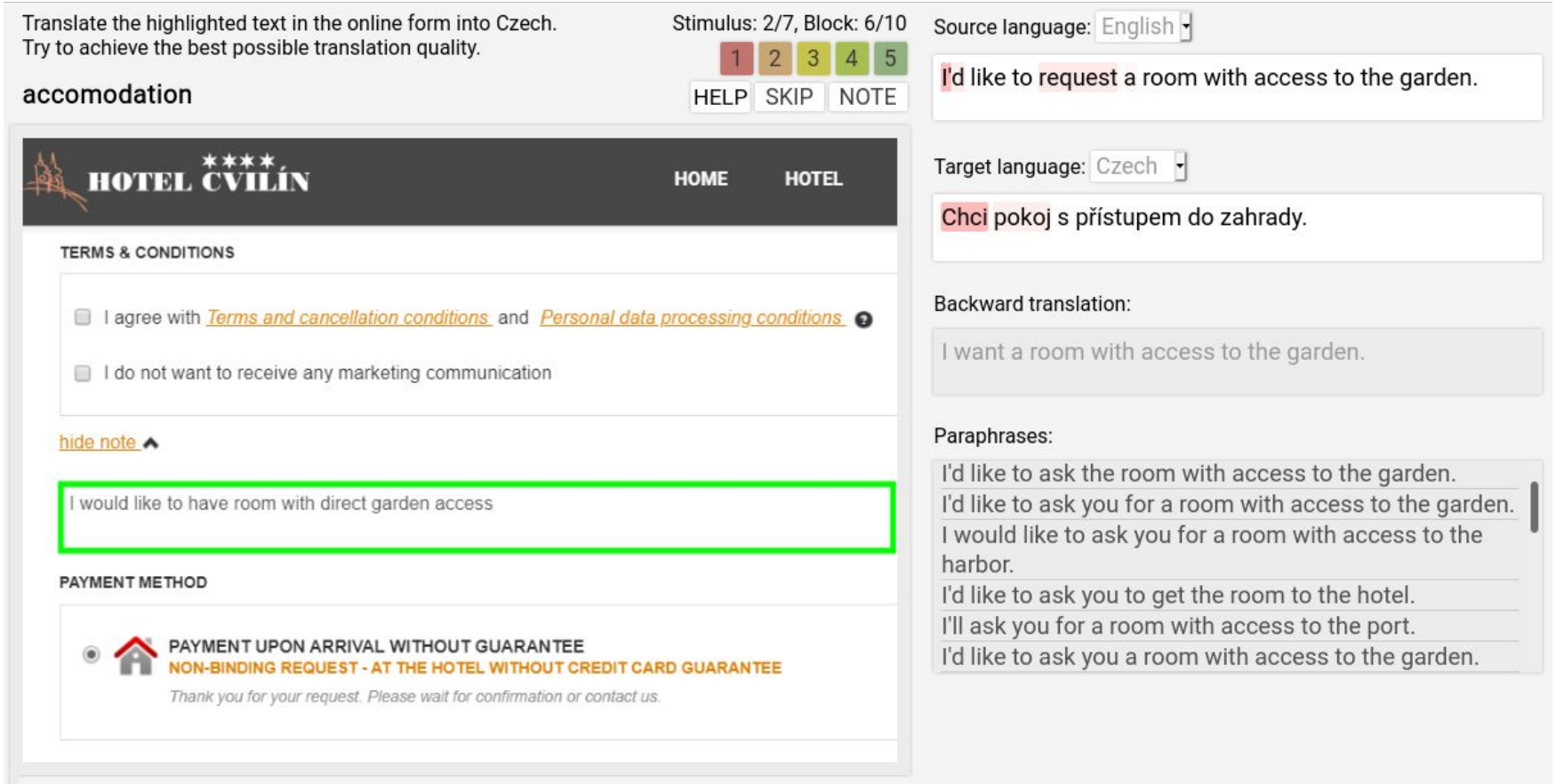
Ptakopět

bergamot

Inbound vs. Outbound Translation



Ptakopět: environment for OT experiments



Ptakopět: environment for OT experiments

- *A user wants to translate to a language that they do not understand*
- Receive cues that should inform about translation quality
- May rephrase the query while keeping its intention

Ptakopět: environment for OT experiments

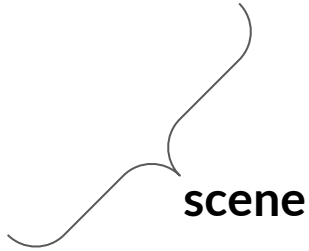
- *A user wants to translate to a language that they do not understand*
- Receive cues that should inform about translation quality
- May rephrase the query while keeping its intention
- **Objectives:**
  - High quality of the translation
  - High confidence of the user about the translation quality

Ptakopět: environment for OT experiments

- *A user wants to translate to a language that they do not understand*
- Receive cues that should inform about translation quality
- May rephrase the query while keeping its intention
- **Objectives:**
  - High quality of the translation
  - High confidence of the user about the translation quality
- **Controlled variables:**
  - Cue modules
  - MT systems (target language, type)

Ptakopět: environment for OT experiments

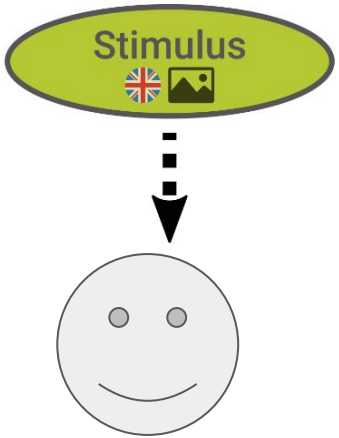
- *A user wants to translate to a language that they do not understand*
- Receive cues that should inform about translation quality
- May rephrase the query while keeping its intention
- **Objectives:**
  - High quality of the translation
  - High confidence of the user about the translation quality
- **Controlled variables:**
  - Cue modules
  - MT systems (target language, type)



Experiment scene

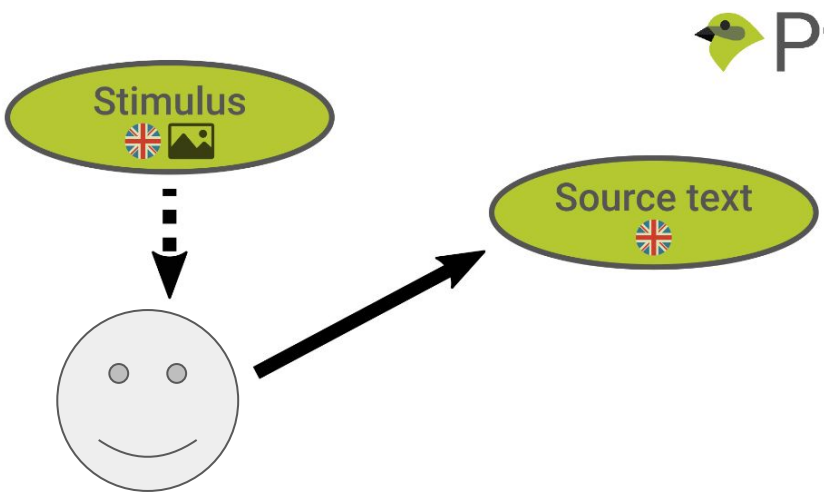
Ptakopět

Experiment scene

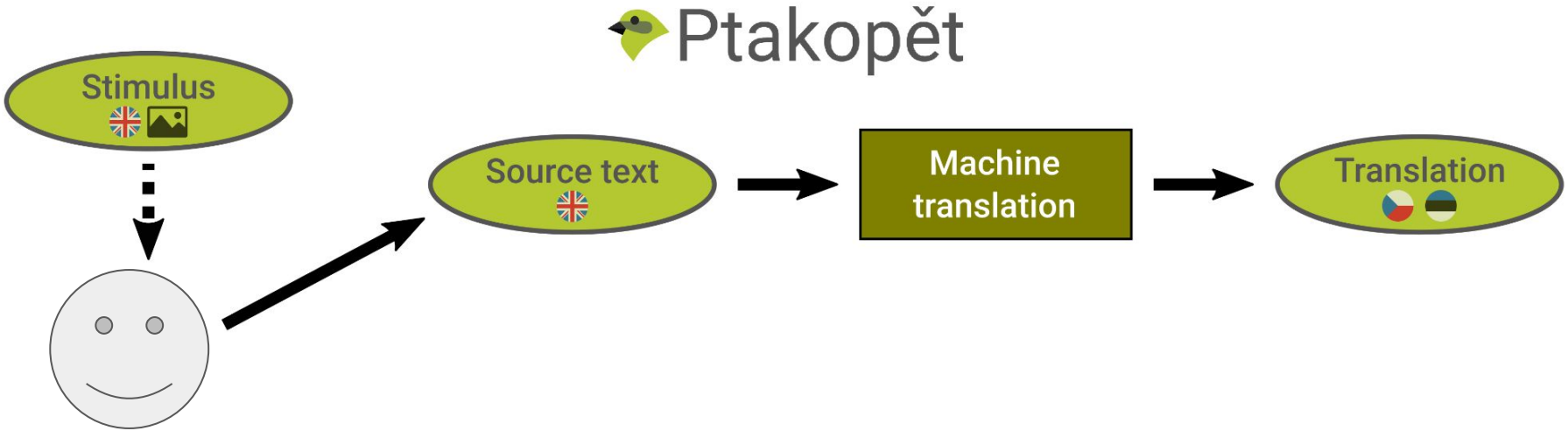


Ptakopět

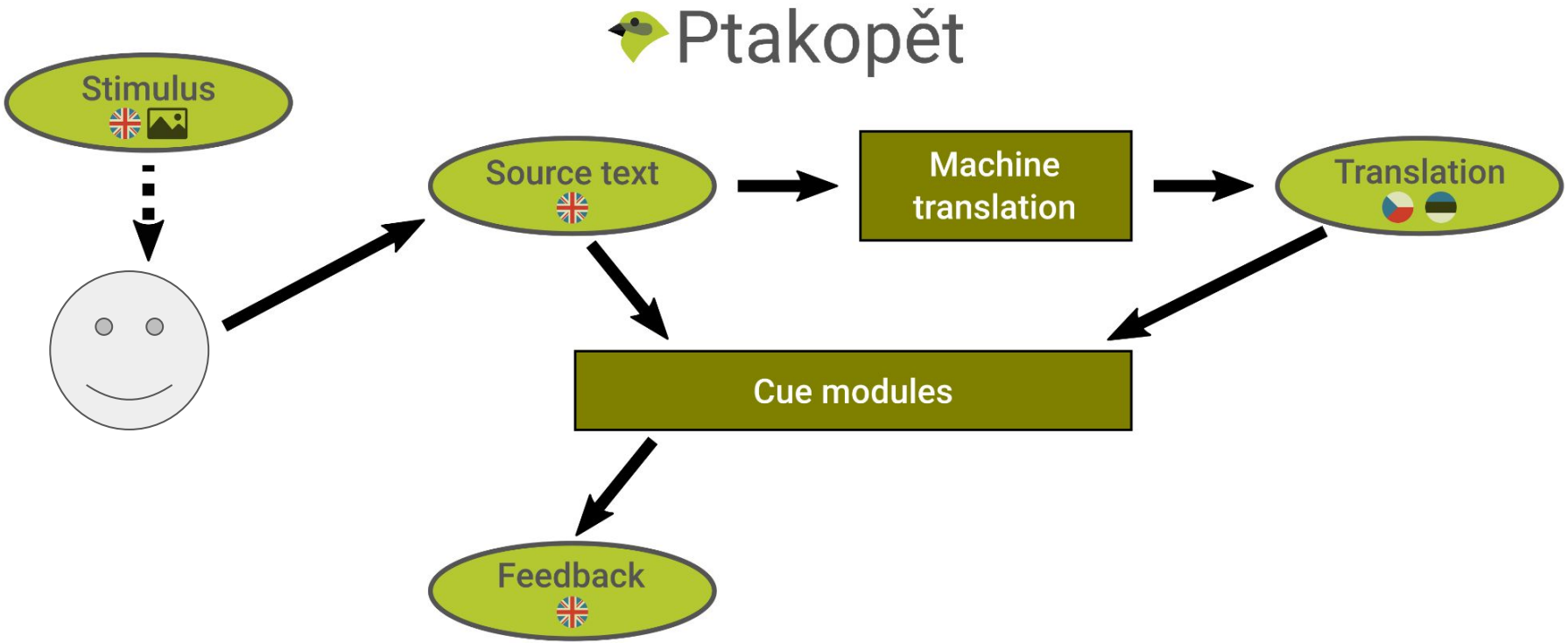
Experiment scene



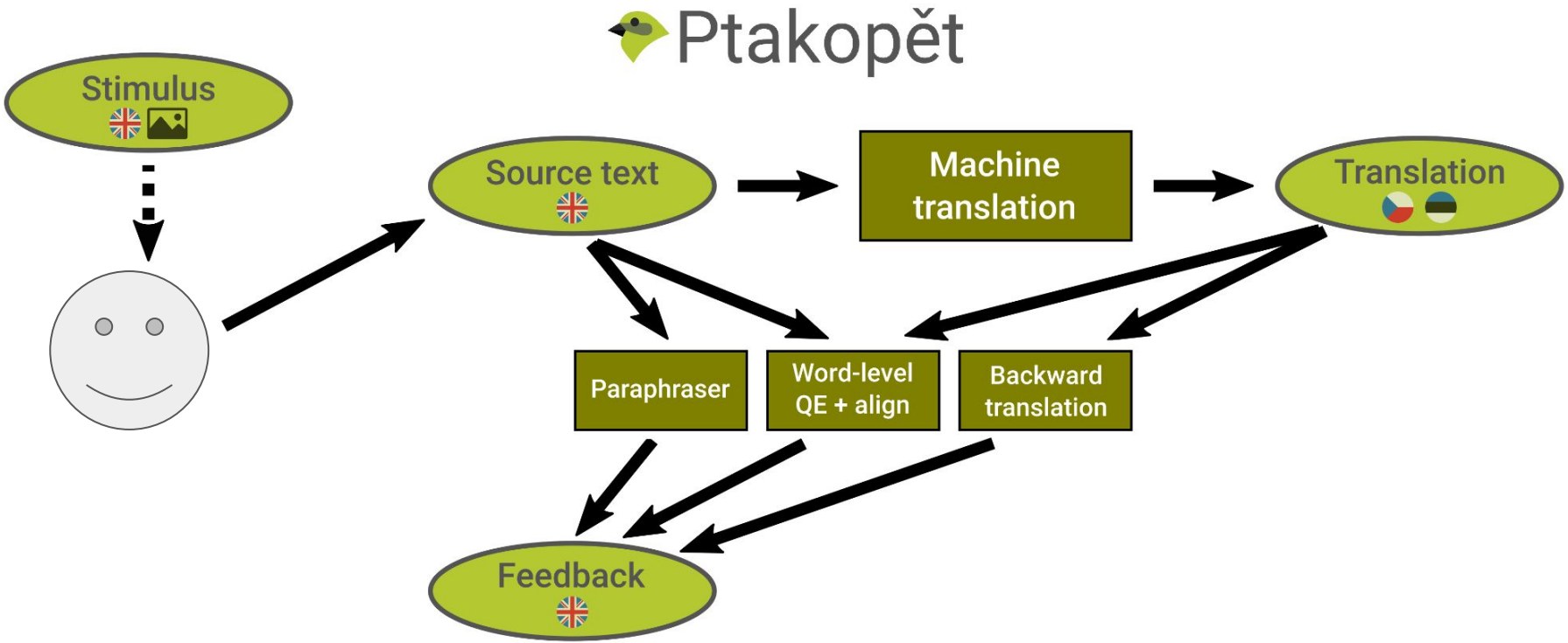
Experiment scene



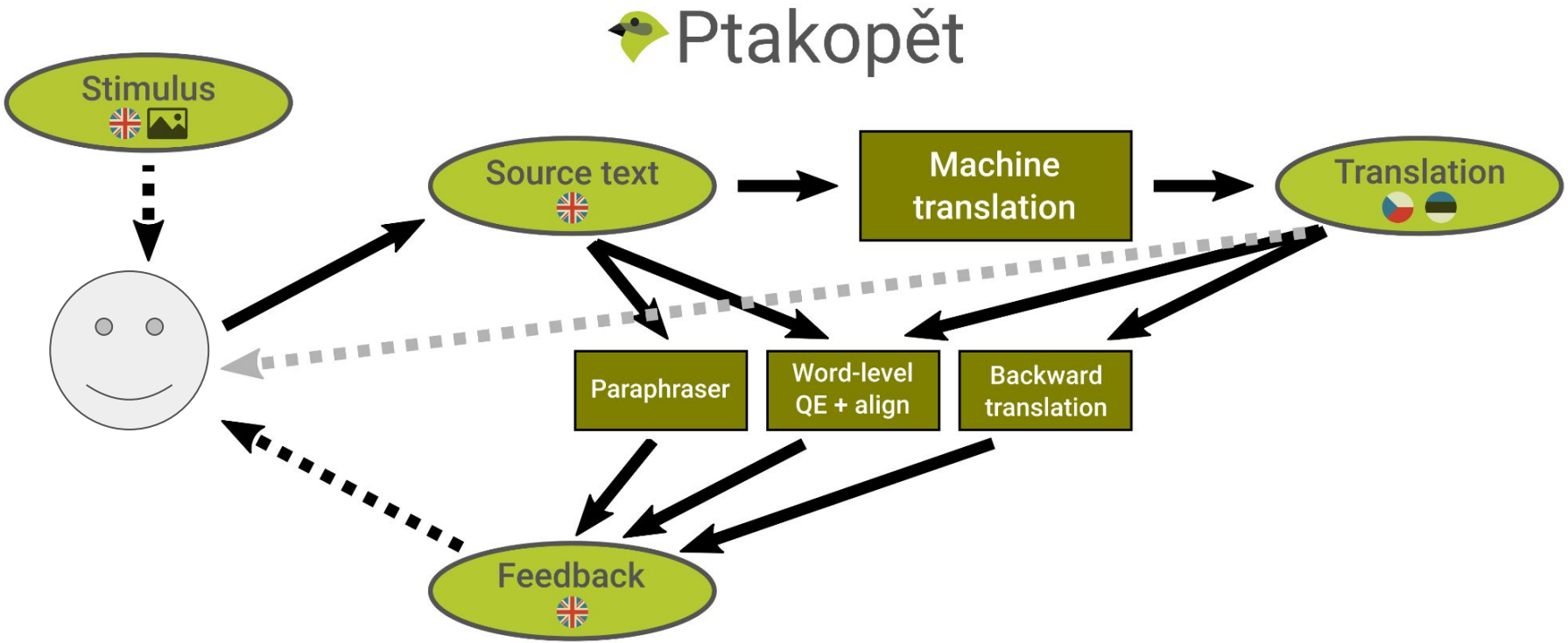
Experiment scene



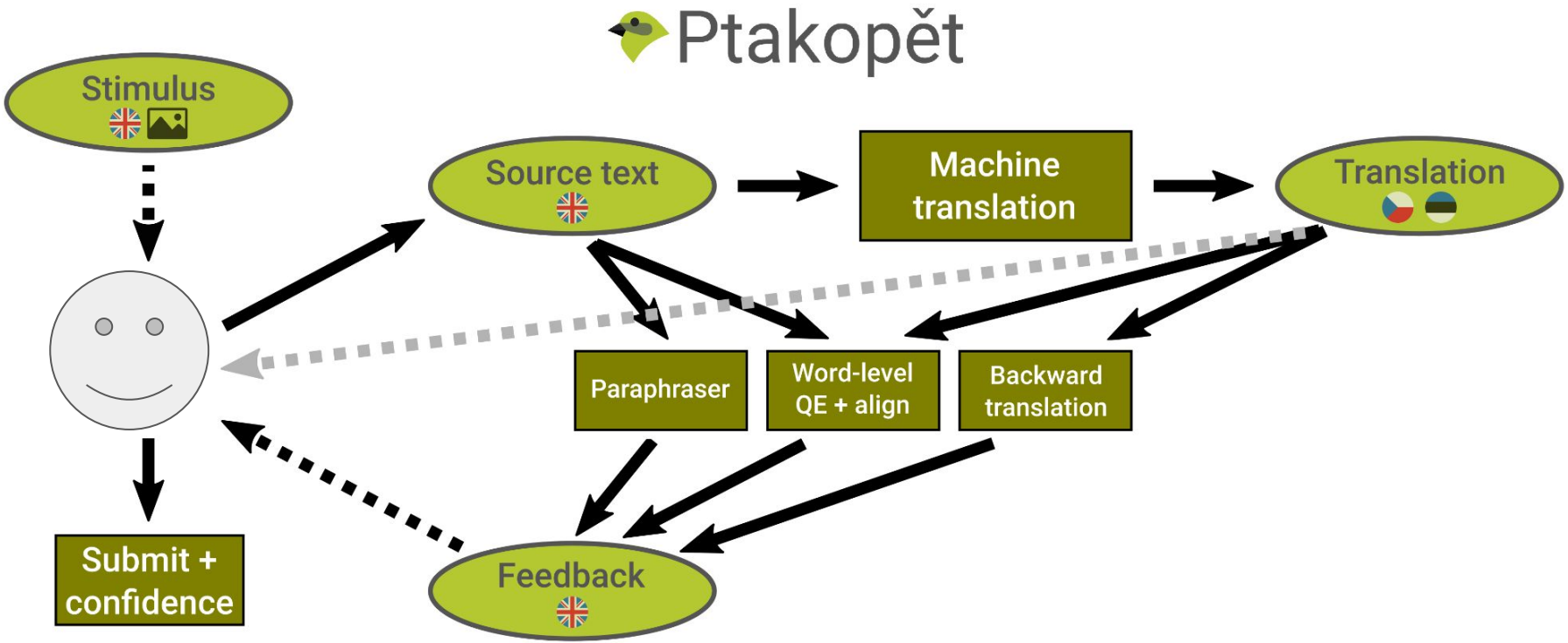
Experiment scene



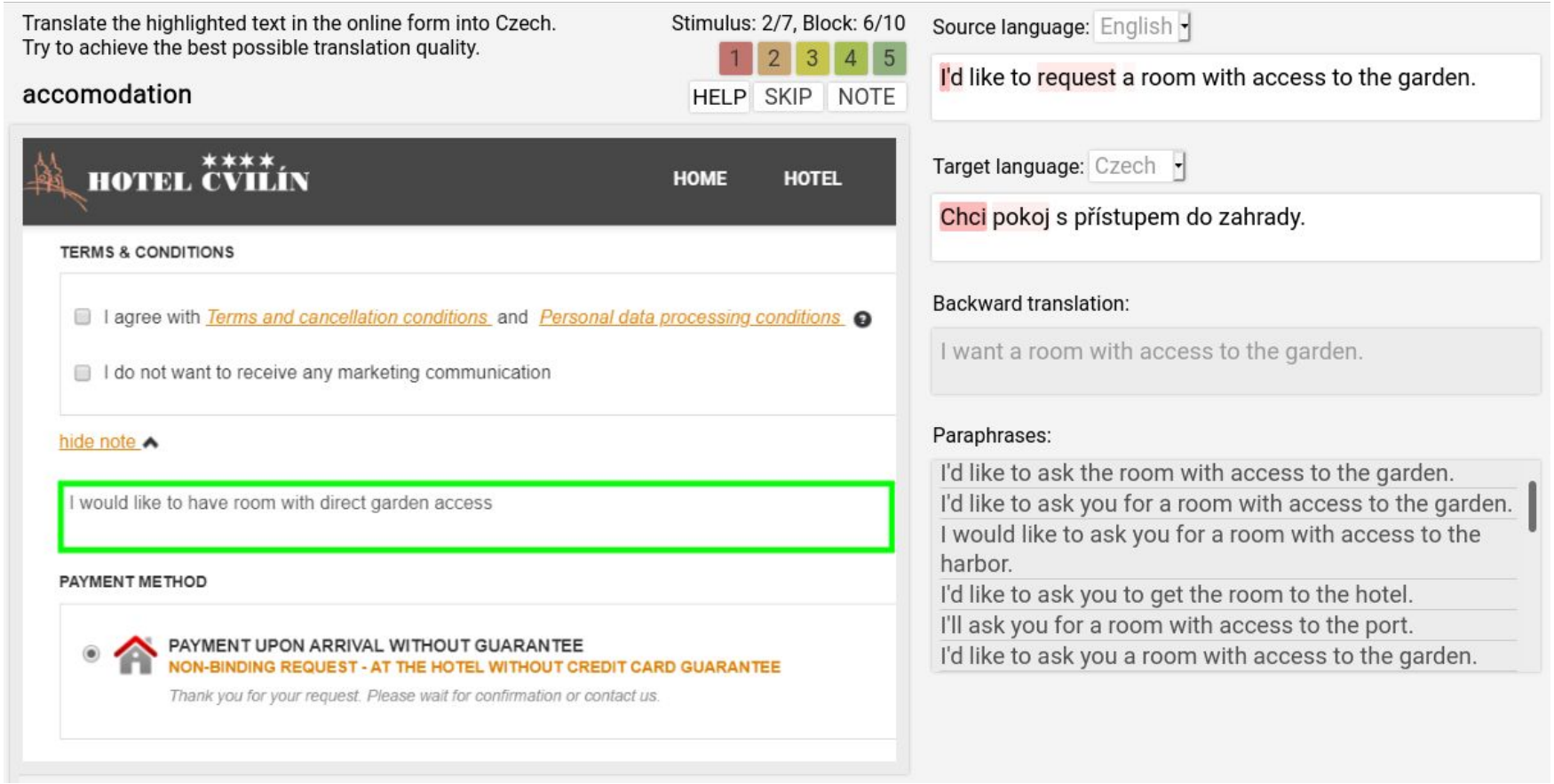
Experiment scene



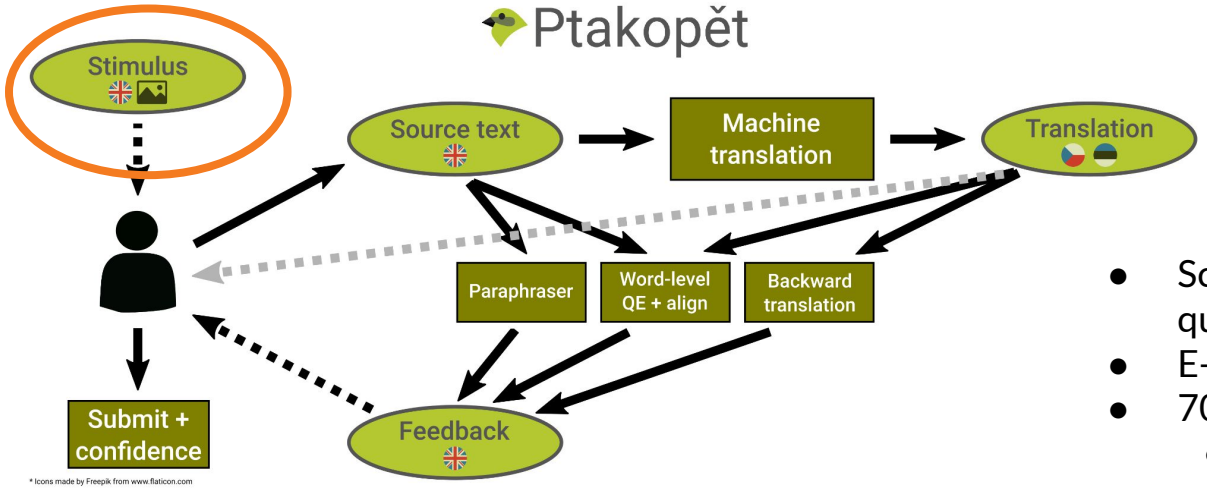
Experiment scene

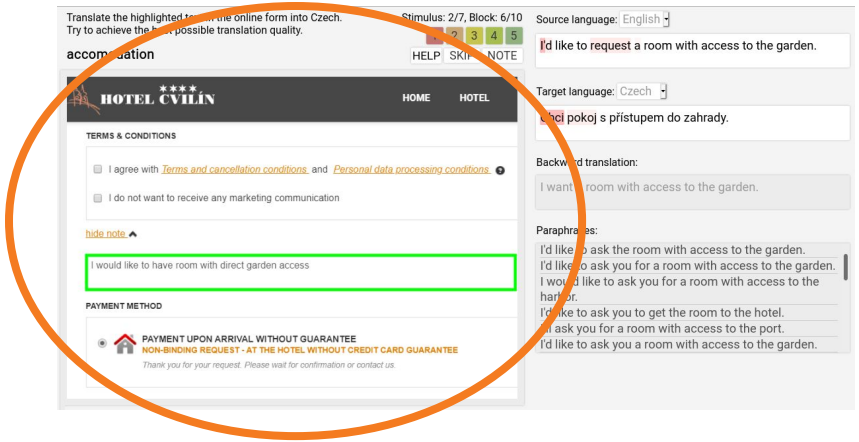


Experiment scene



Stimuli



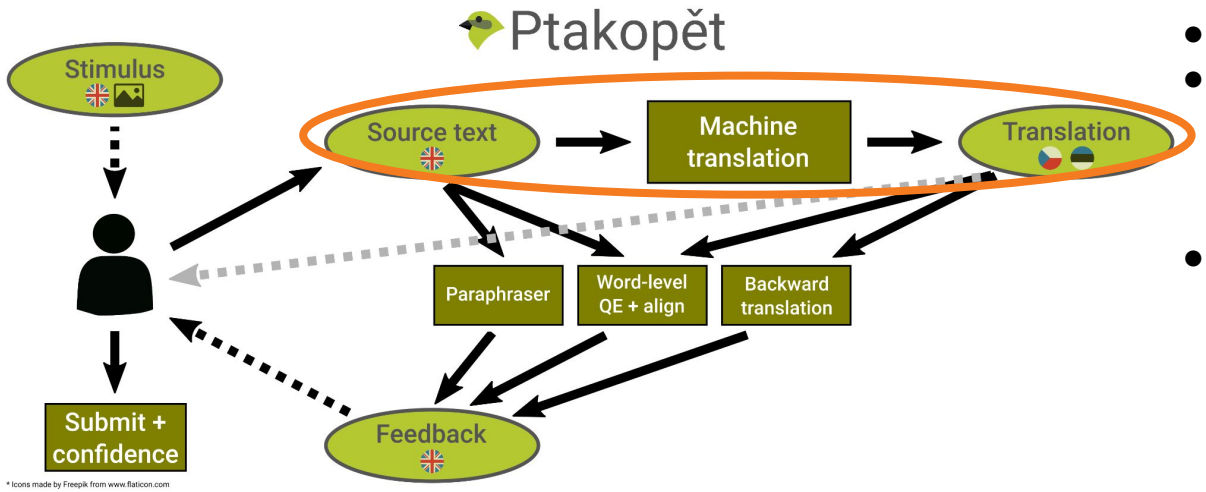


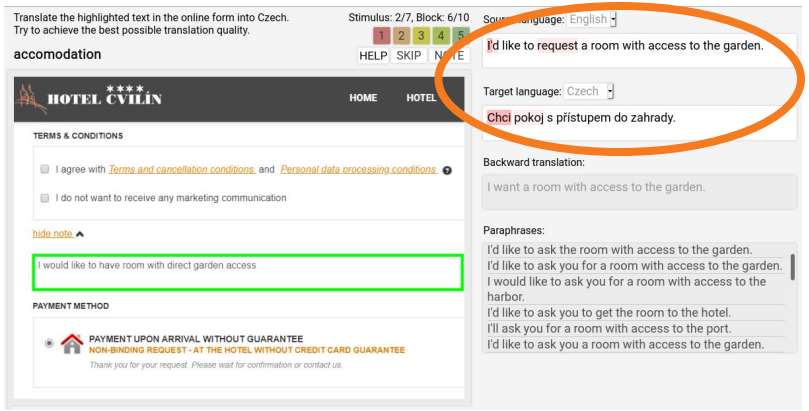
- Screenshots of web forms with a pre-filled query + website topic
- E-commerce domain
- 70 screenshots
  - Collected by 6 annotators
  - Annotators looked for web forms that contain fields to be filled with a text that requires translation

Example

STIM: [accomodation] I would like to have room with direct garden access.

Machine Translation



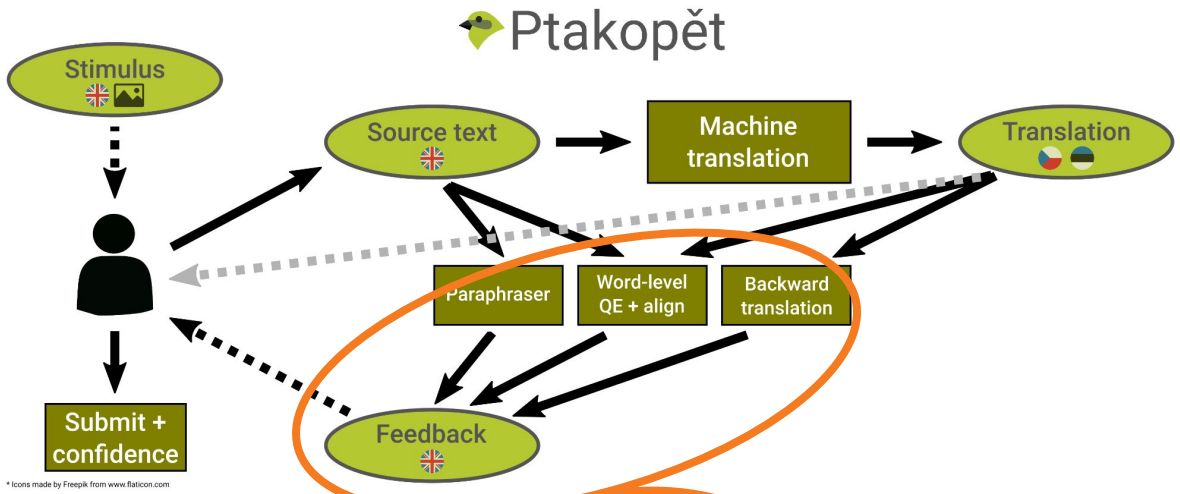


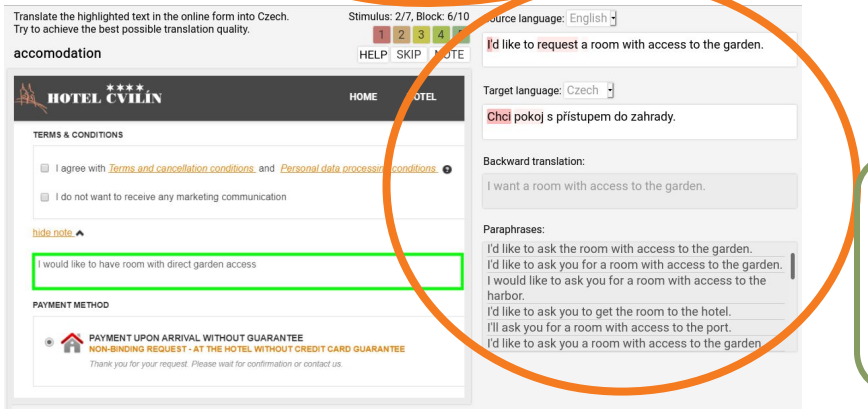
- Transformer-based models
- Czech Teacher (Czech 3)
  - Winning CS-EN system at WMT'19 News Translation Shared Task
- Czech Student (Czech 2)
  - CPU-optimized student created by knowledge distillation using Teacher
  - Much faster
  - Worse by 0.5-3.0 BLEU
- Czech weak teacher (Czech 1)
- Estonian Student
- Randomly select

Example

STIM: I would like to have room with direct garden access.
SRC: I'd like to request a room with access to the garden.
MT: Chci pokoj s přístupem do zahrady.

Cue modules



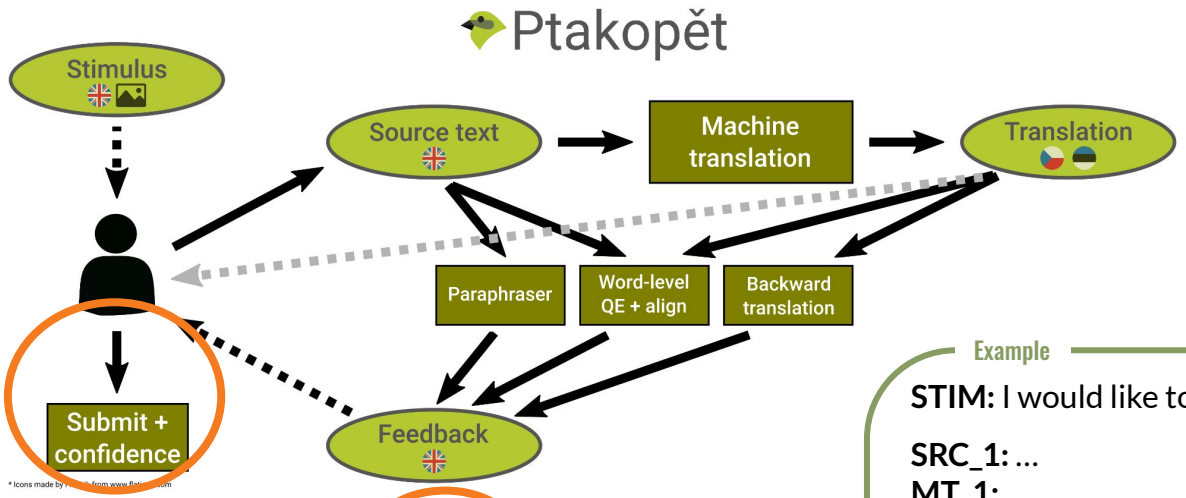


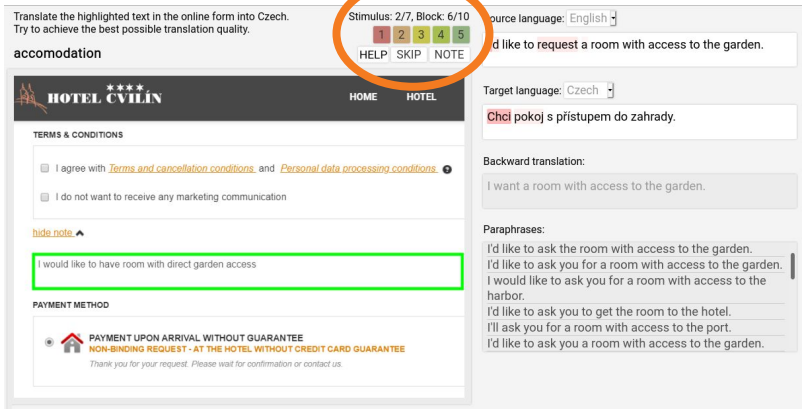
- Backward translation (*BT*)
  - Same MT in reversed direction
- Word-level quality estimation (*QE*)
  - Glass-box features exploiting information from MT system
  - Projected to source by word alignment
  - Only 14-37% F-score of detecting errors
- Paraphrasing (*PP*)
  - Pivot translation via 41 languages
  - Grouped by similarity
- Random combination (even empty)

Example

STIM: I would like to have room with direct garden access.
SRC: I'd like to request a room with access to the garden.
MT: Chci pokoj s přístupem do zahrady.
BT: I want a room with access to the garden.
PARAPH: I'd like to ask for a room with access to the garden.

Self-reported confidence





- How much the user believes the translation is correct
  - from 1 (least) to 5 (most)

Example

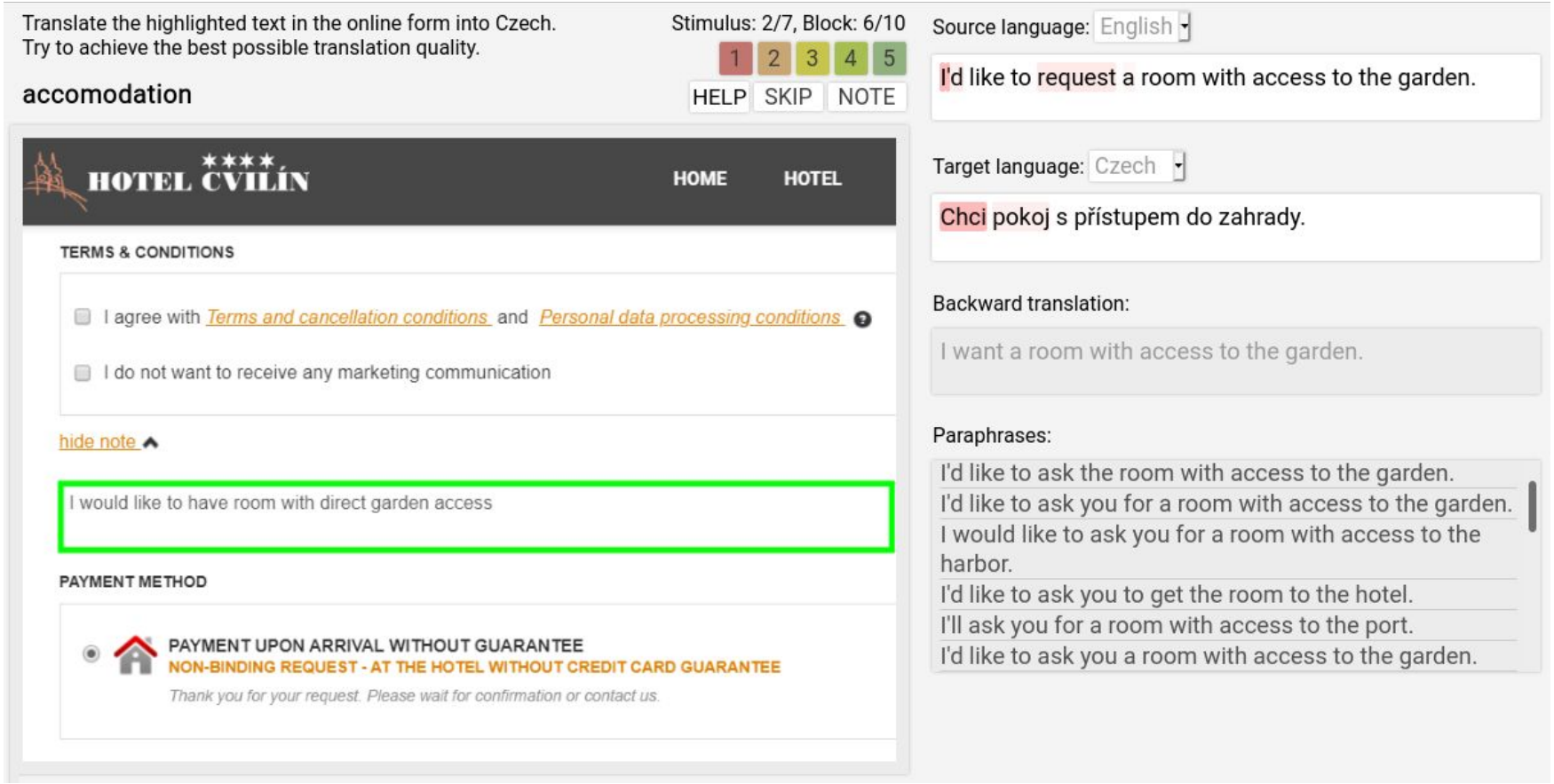
STIM: I would like to have room with direct garden access.

SRC_1: …
MT_1: …
BT_1: …
PARAPH_1: …

…

SRC_n: I'd like to request a room with access to the garden.
MT_n: Chci pokoj s přístupem do zahrady.
BT_n: I want a room with access to the garden.
PARAPH_n: I'd like to ask the room with access to the garden.
CONF: 4

Experiment scene



Experiment details

- 52 English speaking participants (49 native)
- 70 scenes with a random combination of an MT system and cue modules
- 2,486 valid scenes collected in total (48 per user on average)

Experiment details

- 52 English speaking participants (49 native)
- 70 scenes with a random combination of an MT system and cue modules
- 2,486 valid scenes collected in total (48 per user on average)

	# Scenes	Time [s]	Actions	Pace [s]
CS Teacher	601	52	1.66	41
CS Student	643	42	3.66	17
+BT	1253	49	2.55	29
+QE	1240	49	2.86	26
+PP	1211	51	2.75	28

Experiment details

- 52 English speaking participants (49 native)
- 70 scenes with a random combination of an MT system and cue modules
- 2,486 valid scenes collected in total (48 per user on average)

	# Scenes	Time [s]	Actions	Pace [s]
CS Teacher	601	52	1.66	41
CS Student	643	42	3.66	17
+BT	1253	49	2.55	29
+QE	1240	49	2.86	26
+PP	1211	51	2.75	28

Experiment details

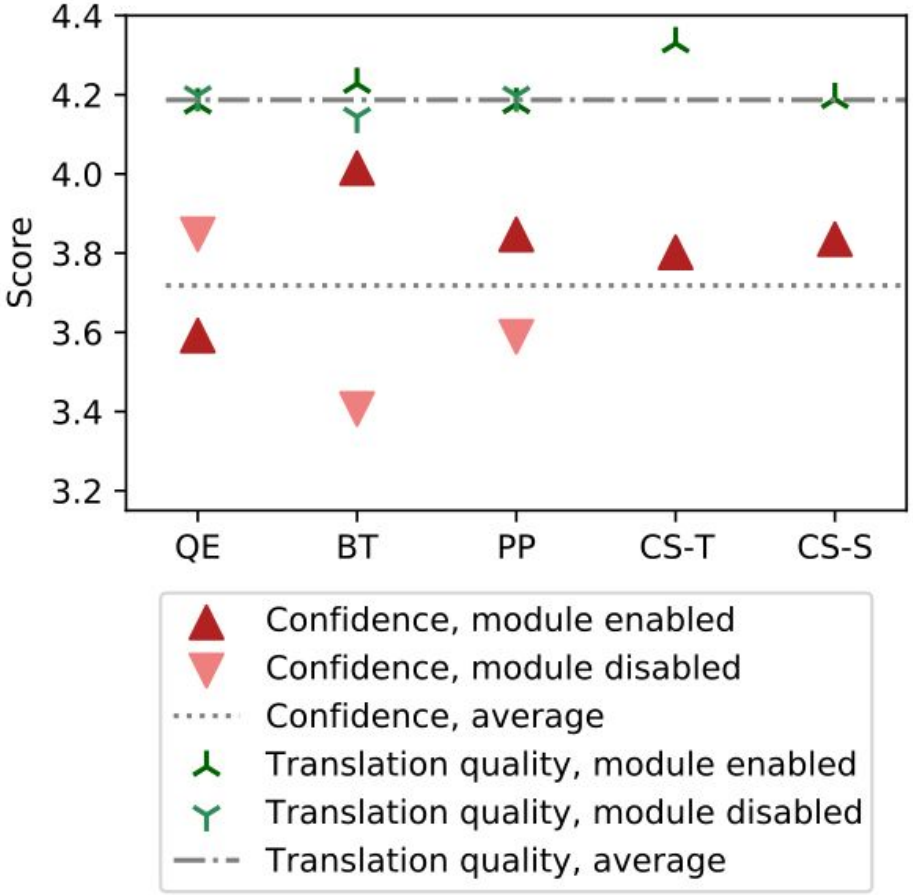
- 52 English speaking participants (49 native)
- 70 scenes with a random combination of an MT system and cue modules
- 2,486 valid scenes collected in total (48 per user on average)

	# Scenes	Time [s]	Actions	Pace [s]
CS Teacher	601	52	1.66	41
CS Student	643	42	3.66	17
+BT	1253	49	2.55	29
+QE	1240	49	2.86	26
+PP	1211	51	2.75	28

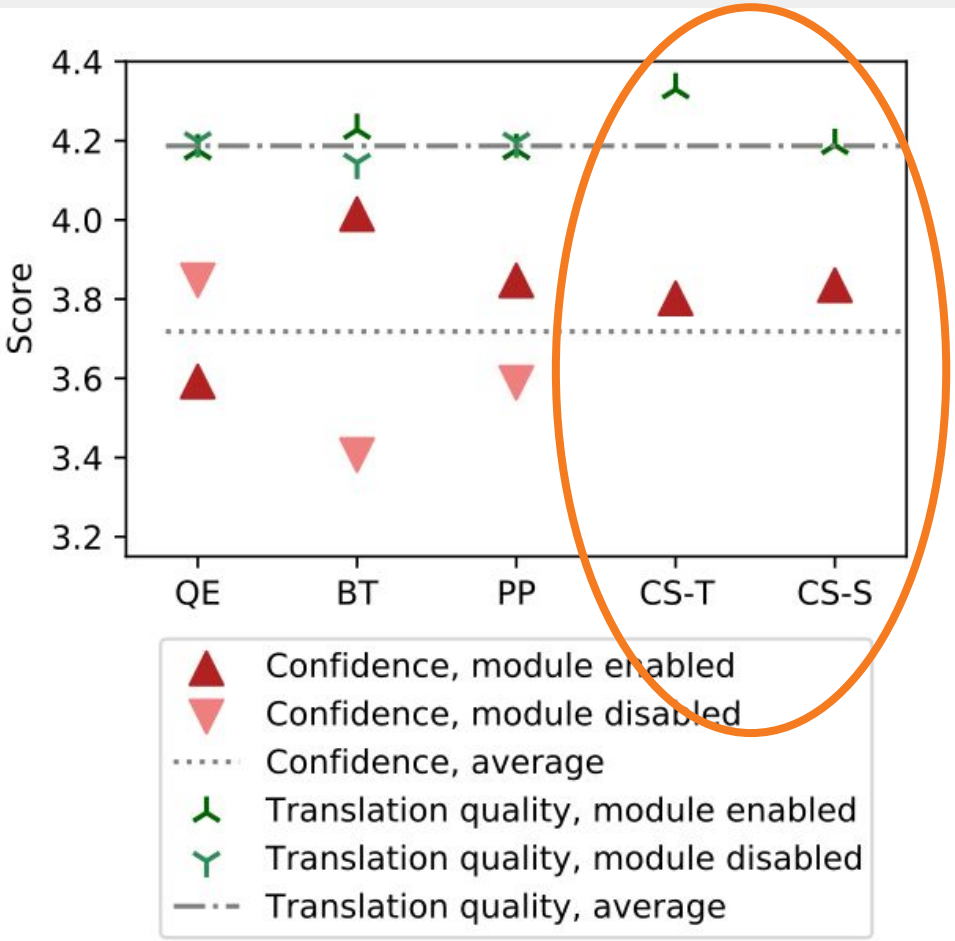
MT evaluation

- Human evaluation by 12 Czech and 3 Estonian native speakers
- Based on:
  - Source (SRC), its translation (TGT) and the corresponding stimulus (STI)
- Rating from 1 (least) to 5 (most) = the scale of self-reported confidence:
  - Overall quality
  - SRC-STI adequacy
  - TGT-SRC adequacy
  - TGT-STI adequacy
  - Fluency

Results

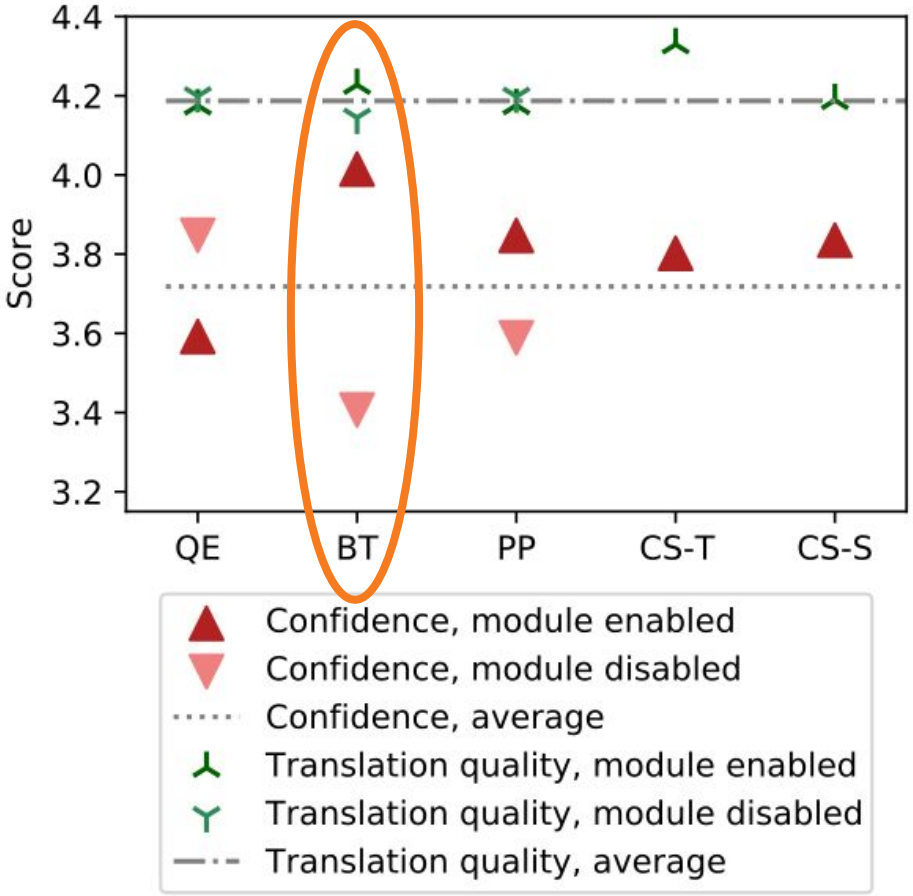

4.4
4.2
4.0
3.8
3.6
3.4
3.2
Score
QE
BT
PP
CS-T
CS-S
Confidence, module enabled
Confidence, module disabled
Confidence, average
Translation quality, module enabled
Translation quality, module disabled
Translation quality, average

Results



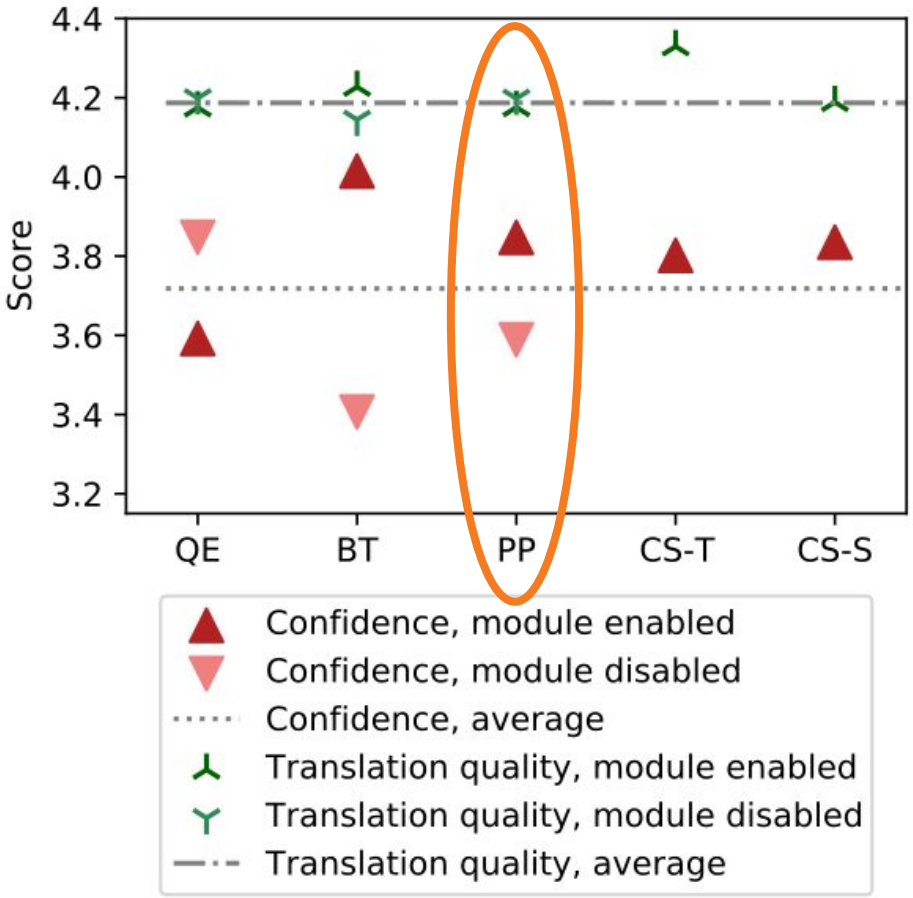
- Despite the higher quality of CS Teacher, users are slightly more confident with CS Student
  - Possibly high Student speed

Results



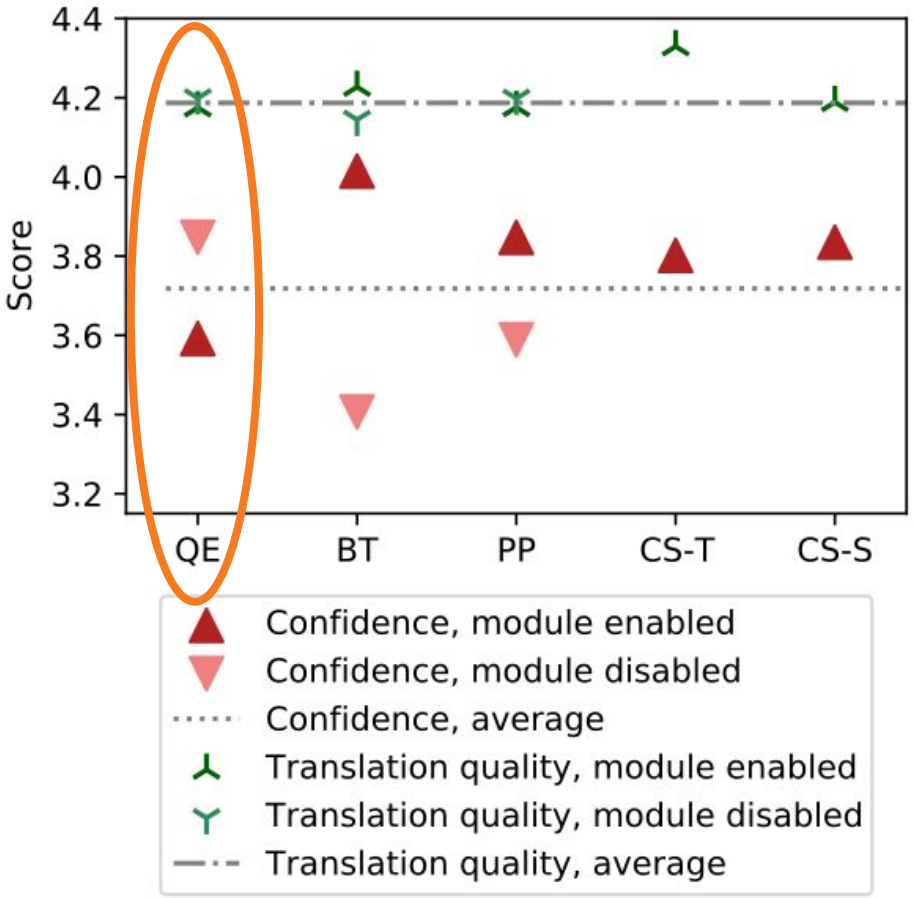
- Despite the higher quality of CS Teacher, users are slightly more confident with CS Student
  - Possibly high Student speed
- BT feedback significantly increases confidence with little impact on quality
  - Great for reassuring users

Results



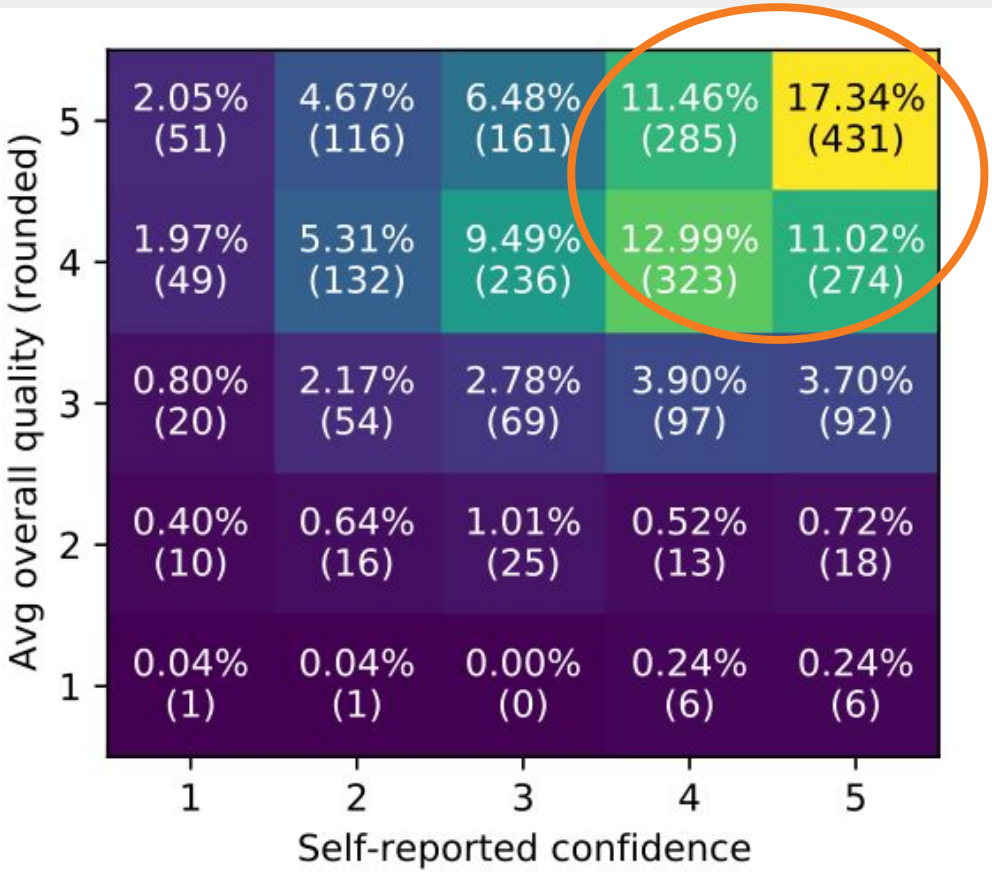
- Despite the higher quality of CS Teacher, users are slightly more confident with CS Student
  - Possibly high Student speed
- BT feedback significantly increases confidence with little impact on quality
  - Great for reassuring users
- PP insignificantly increases confidence
  - Practical use is questionable

Results



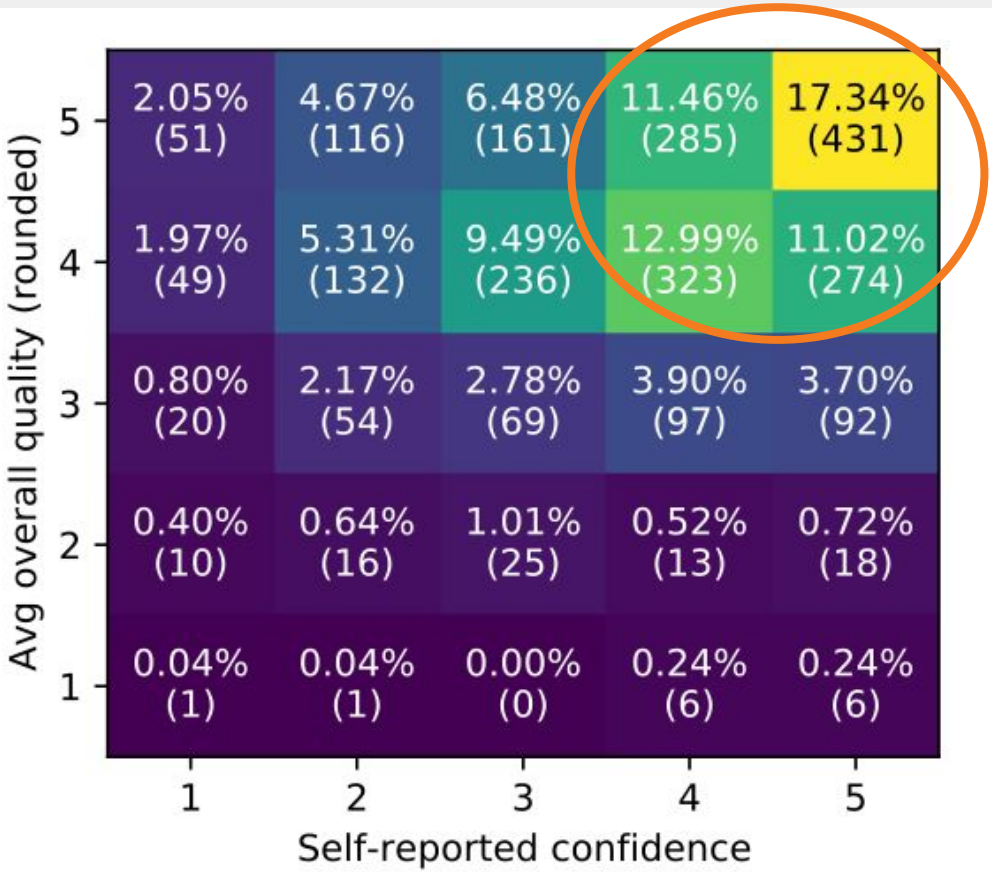
- Despite the higher quality of CS Teacher, users are slightly more confident with CS Student
  - Possibly high Student speed
- BT feedback significantly increases confidence with little impact on quality
  - Great for reassuring users
- PP insignificantly increases confidence
  - Practical use is questionable
- QE decreases confidence
  - Word-level models not mature enough for user-facing applications

Results



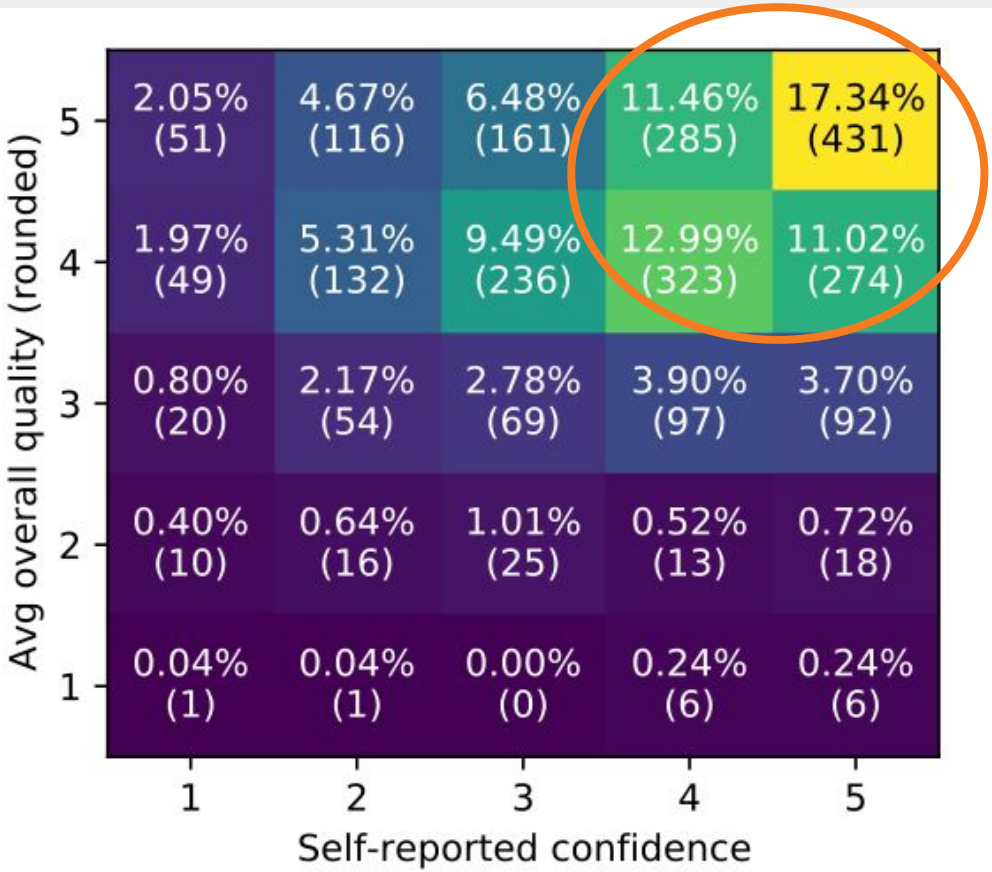
- Both confidence and quality scores cluster around the highest values
  - 63% of scenes received 4 or 5 in both scores

Results



- Both confidence and quality scores cluster around the highest values
  - 63% of scenes received 4 or 5 in both scores
- Positive in terms of industry progress, yet harder to analyze

Results



- Both confidence and quality scores cluster around the highest values
  - 63% of scenes received 4 or 5 in both scores
- Positive in terms of industry progress, yet harder to analyze
- Could be mitigated by:
  - Much lower-resourced language pairs
  - More complex queries to translate

Examples

Inter	I teach my son **English** with the 'Learning Time with Timmy' series on Youtube.
	Učím svého syna **Angličana** */Englishman/* se seriálem „Learning Time with Timmy" na Youtube.
Final	I teach my son **English language** with the series 'Learning Time with Timmy' series on Youtube.
	Učím svého syna **anglický jazyk** se seriálem „Learning Time with Timmy" na Youtube.
Inter	Why was I not able to make a payment **by** mobile?
	Proč jsem nemohl zaplatit **za** */for/* mobil?
Final	Why was I not able to make a payment **from** my mobile?
	Proč jsem nemohl zaplatit **z** mobilu?
Inter	What documents do I need to have if my ID **has expired**?
	Jaké doklady potřebuji, když mi **vypršel** průkaz totožnosti?
Final	What documents do I need to have if my ID is **out of date**?
	Jaké dokumenty potřebuji, když je můj průkaz **zastaral** */got obsolete/*?

Conclusion

- Inbound vs. Outbound MT problem definition
- Backtranslation helps in confidence
- MT system speed is important even in comparison to objective quality