Language in Interaction

Speaker and Listener information

SS16 - (Embodied) Language Comprehension

Maria Staudte

So far …

- Embodiment
- Situated & embodied language learning
- Situated adult language comprehension (& production)
- Language in Interaction
    - Taking another person into account
    - Sending and perceiving bodily signals

Language in/for Interaction

- Presupposes a listening/speaking partner
- Both partners use more than spoken language
  - Non-verbal signals: Facial expression, emotions, gaze, posture, gesture etc
- How do these influence language processes?
  - Information contribution, timing, cost

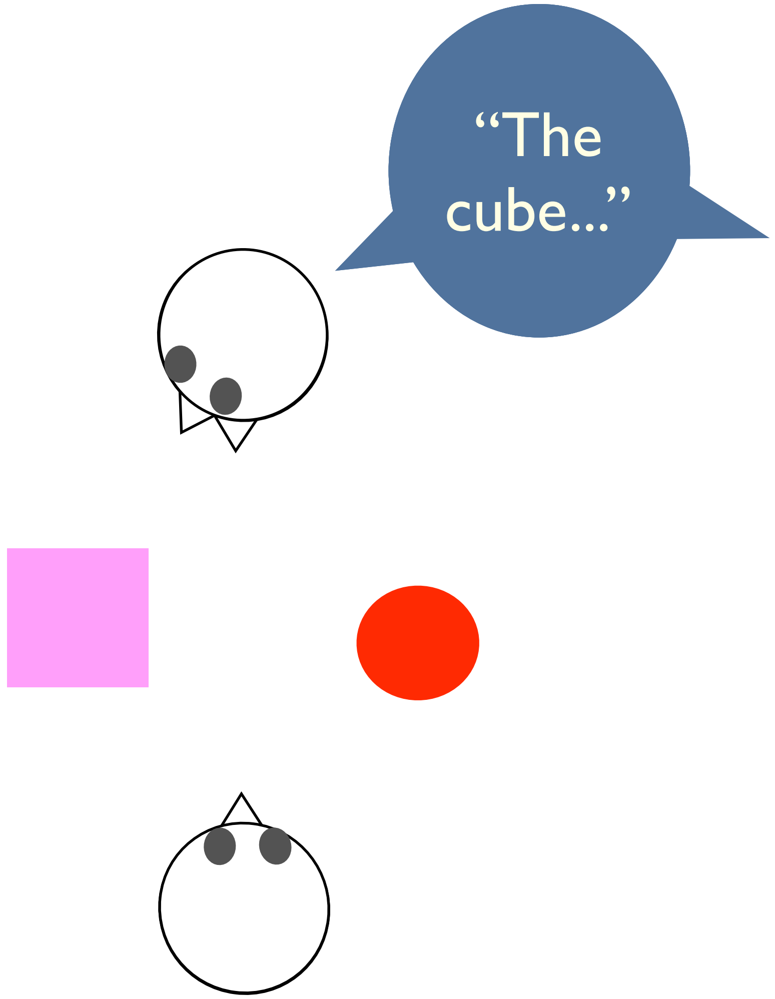
Gaze

- Eye-movements reflect comprehension/prediction/ planning processes
  - Measure
- Eye-movements are a signal by themselves to the partner!
  - Speaker gaze
  - Listener gaze

Speaker gaze

Referential Gaze in Communication

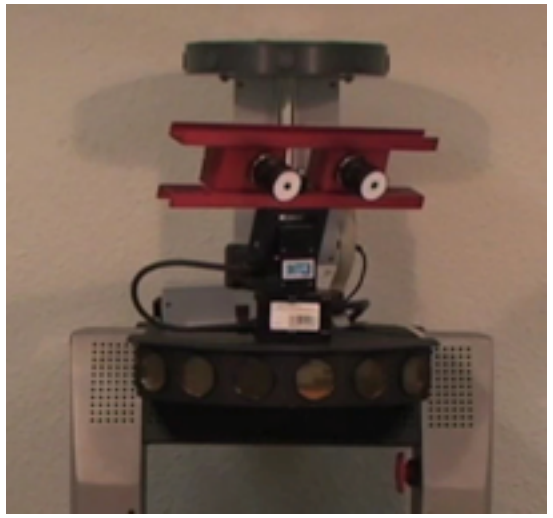
- Speakers look at what they are about to mention (e.g. Griffin & Bock 2000)
- Listeners look at what they hear (e.g. Tanenhaus et al., 1995)
- Listeners look at what the speaker looks at (e.g. Hanna & Brennan 2007)



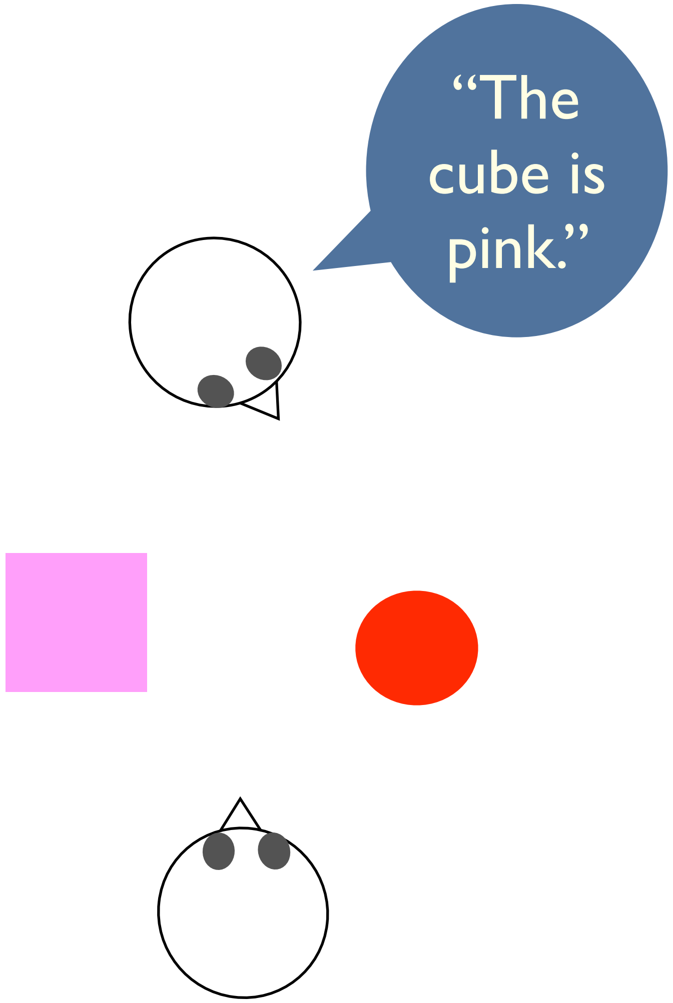
Jim Botsacos in “Cooking together”

(Staudte & Crocker, 2009, 2010, 2011)

Speaker Gaze

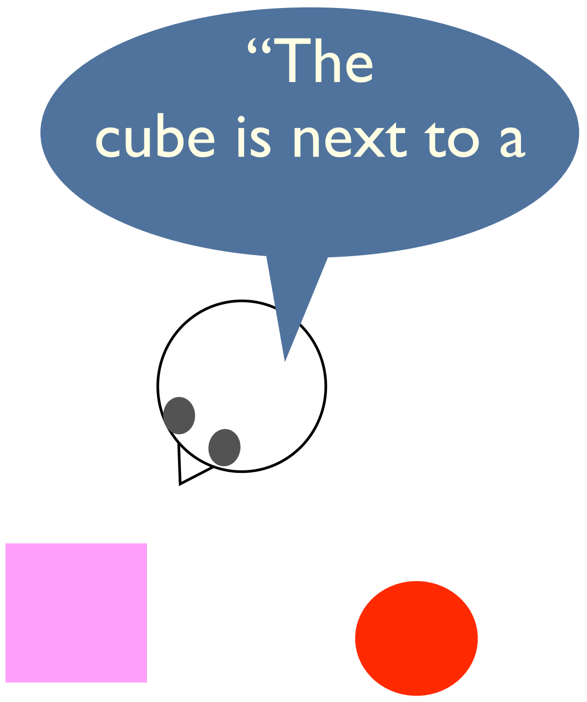


- Listeners follow speaker gaze & utterance
- Facilitation/Disruption effect on sentence validation (congruent vs incongruent)
- Temporal shifts are irrelevant
- Cause of these effects?



Visual Attention & Order

- Speaker gaze & utterance both provide cues that drive listeners' visual attention
- Is order relevant for the utility of cues?
- Manipulate cue order to:
  - Explore information integration
  - Shed light on the role of information provided by speaker gaze



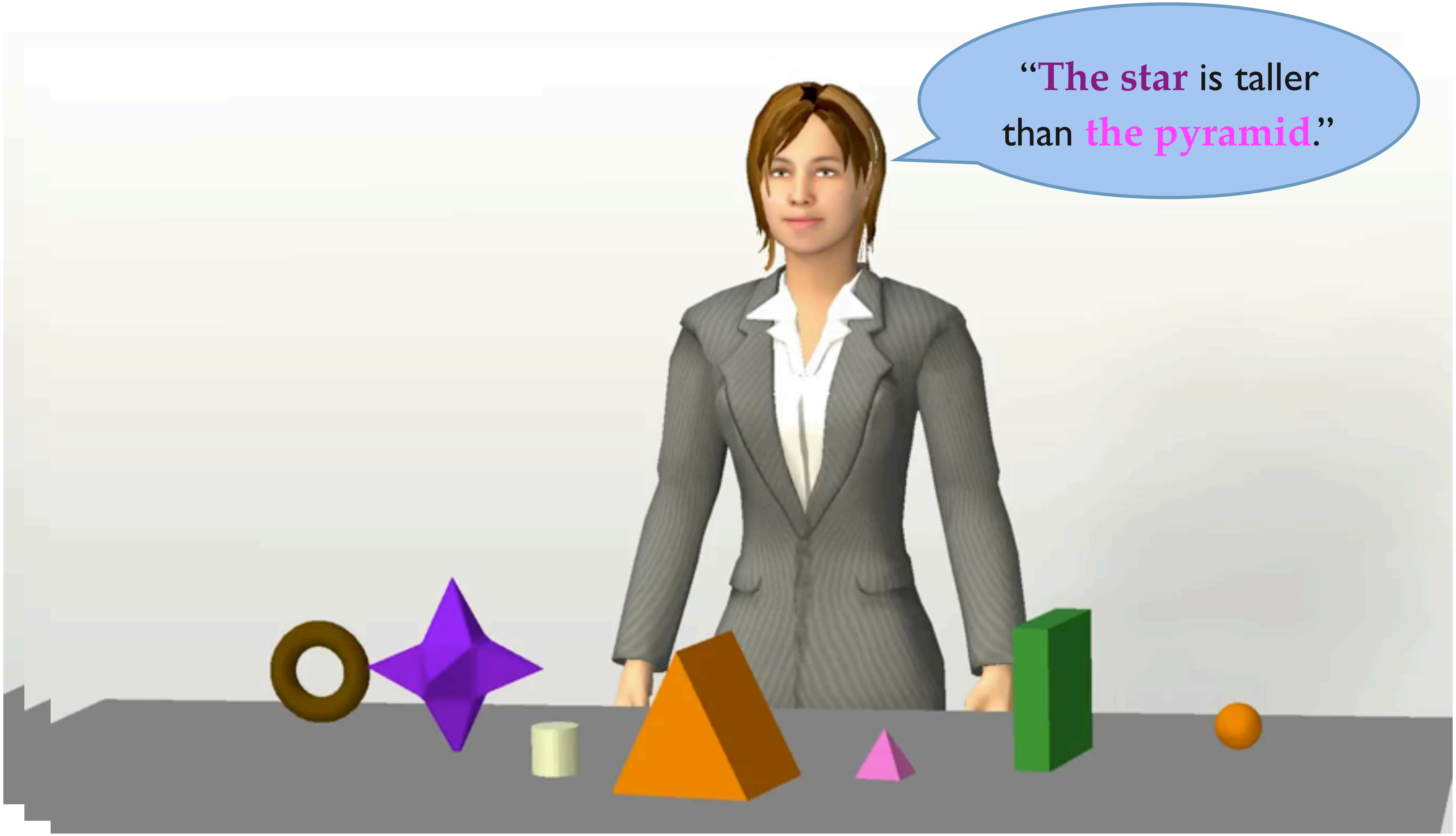
Visual Attention & Order

- Speaker gaze & utterance both provide cues that drive listeners’ visual attention
- Is order relevant for the utility of cues?
- Manipulate cue order to:
    - Is there a bias towards preferring one modality?
    - Shed light on the role of information provided by speaker gaze

Visual Attention & Order

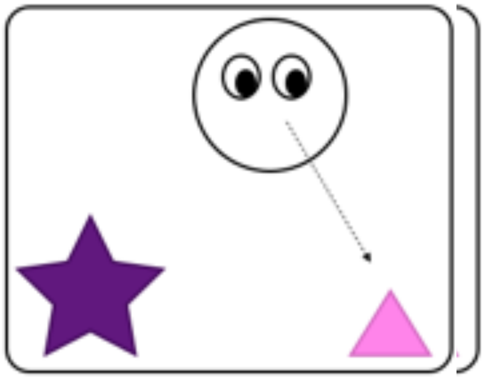
- Speaker gaze & utterance both provide cues that drive listeners' visual attention
- Is order relevant for the utility of cues?
- Manipulate cue order to:
    - Is there a bias towards preferring one modality?
    - Is gaze like any other visual cue, simply increasing visual saliency? (persistent highlighting)
    - Does gaze signal intentions linked to utterance?

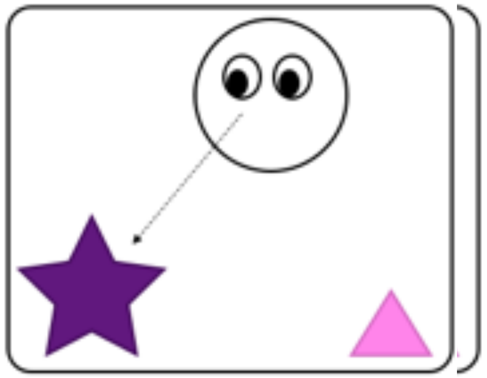
Experiment 1



Experiment 1

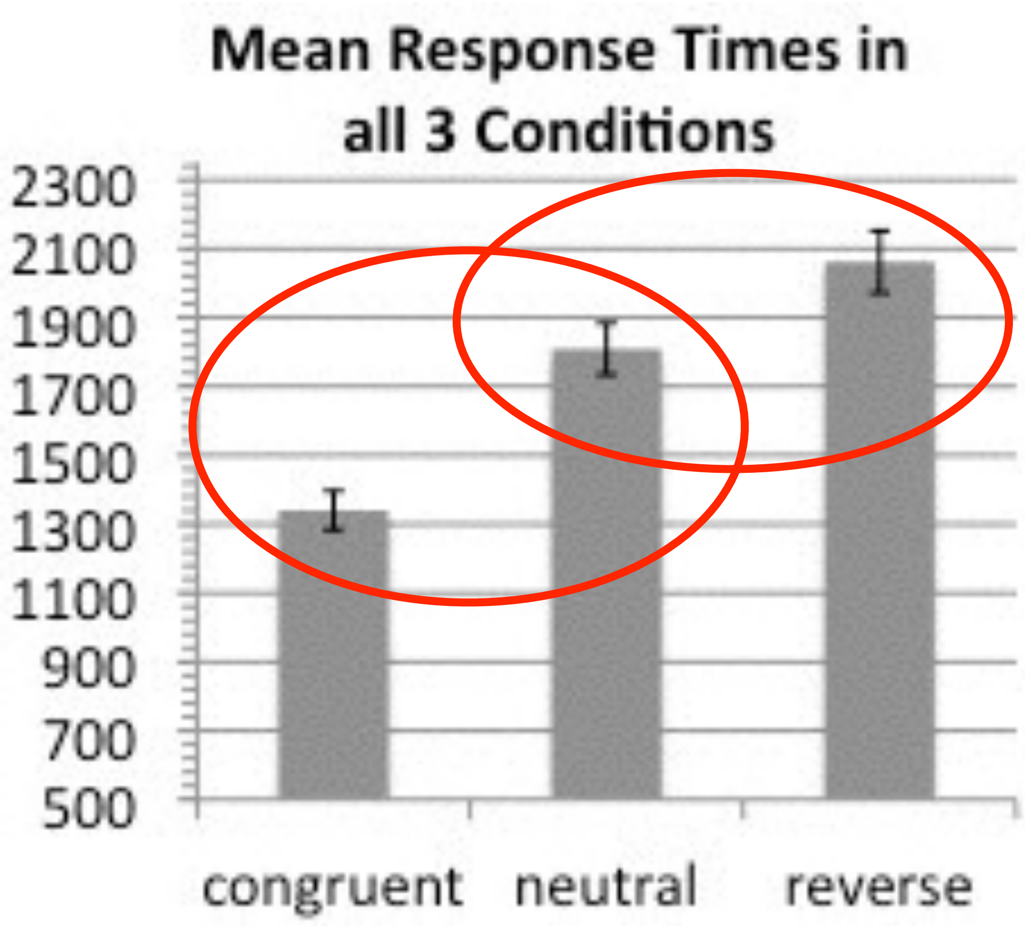
- Task: Is the utterance correct or not?
- 3 Conditions:
  - Congruent, Reverse, Neutral
- 36 diff ts (12 shap
- All co ed
- DV: Eye-movements, Response time

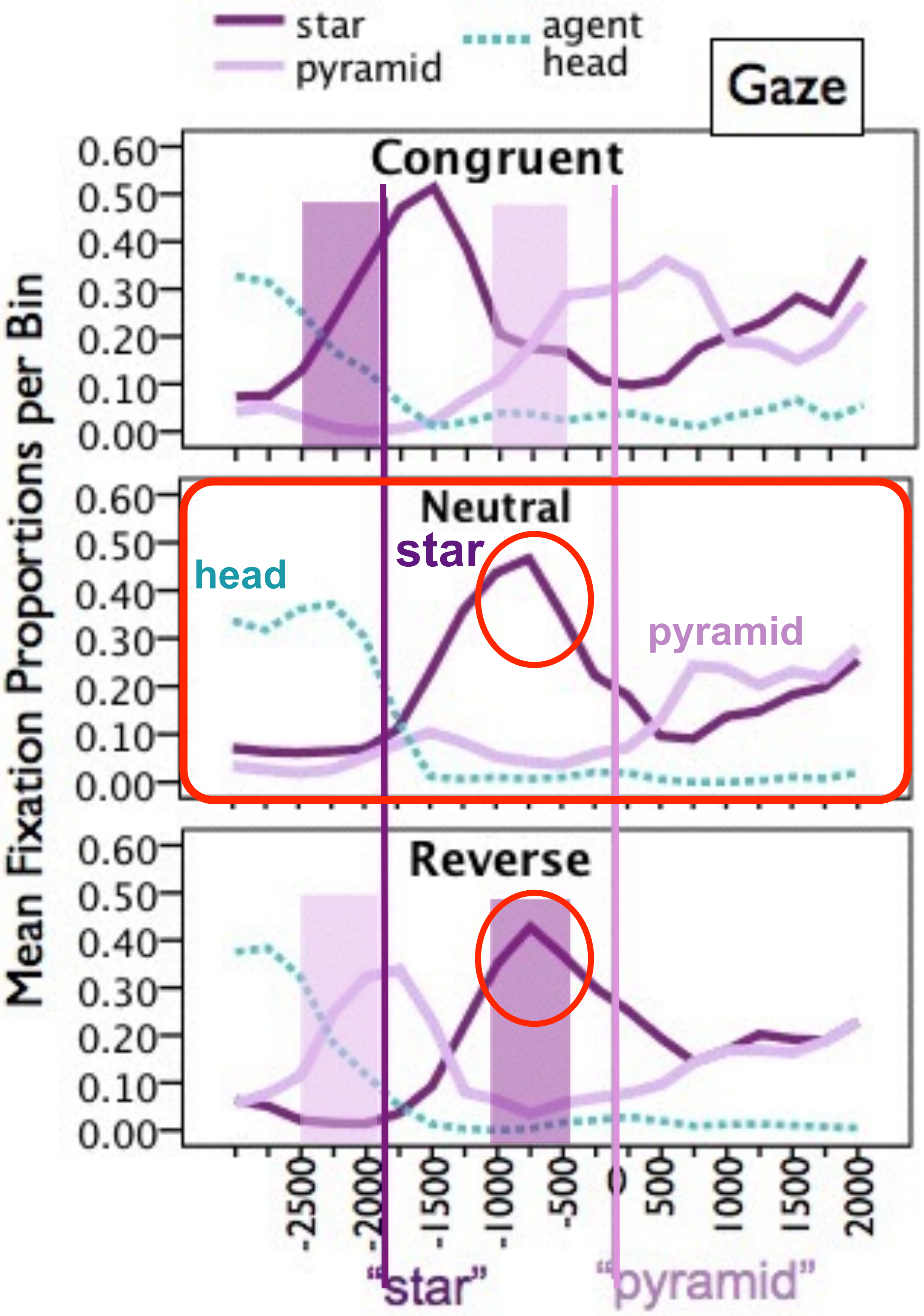




Congruent: Reverse: *The **star** is taller than the pyramid*

Experiment 1 : Results

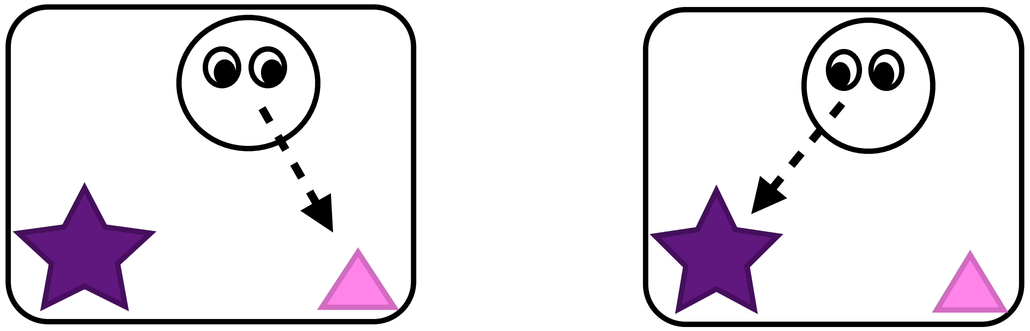




Experiment 1 : Results

- Eye-movement data:
  - Visual attention shifts are elicited by both speaker gaze and utterance, possibly automatically
- Response time data:
  - Visual information, gained through speech- and gaze-mediated attention shifts, is integrated
  - Information integration is difficult in reverse condition!

Experiment 1 : Results

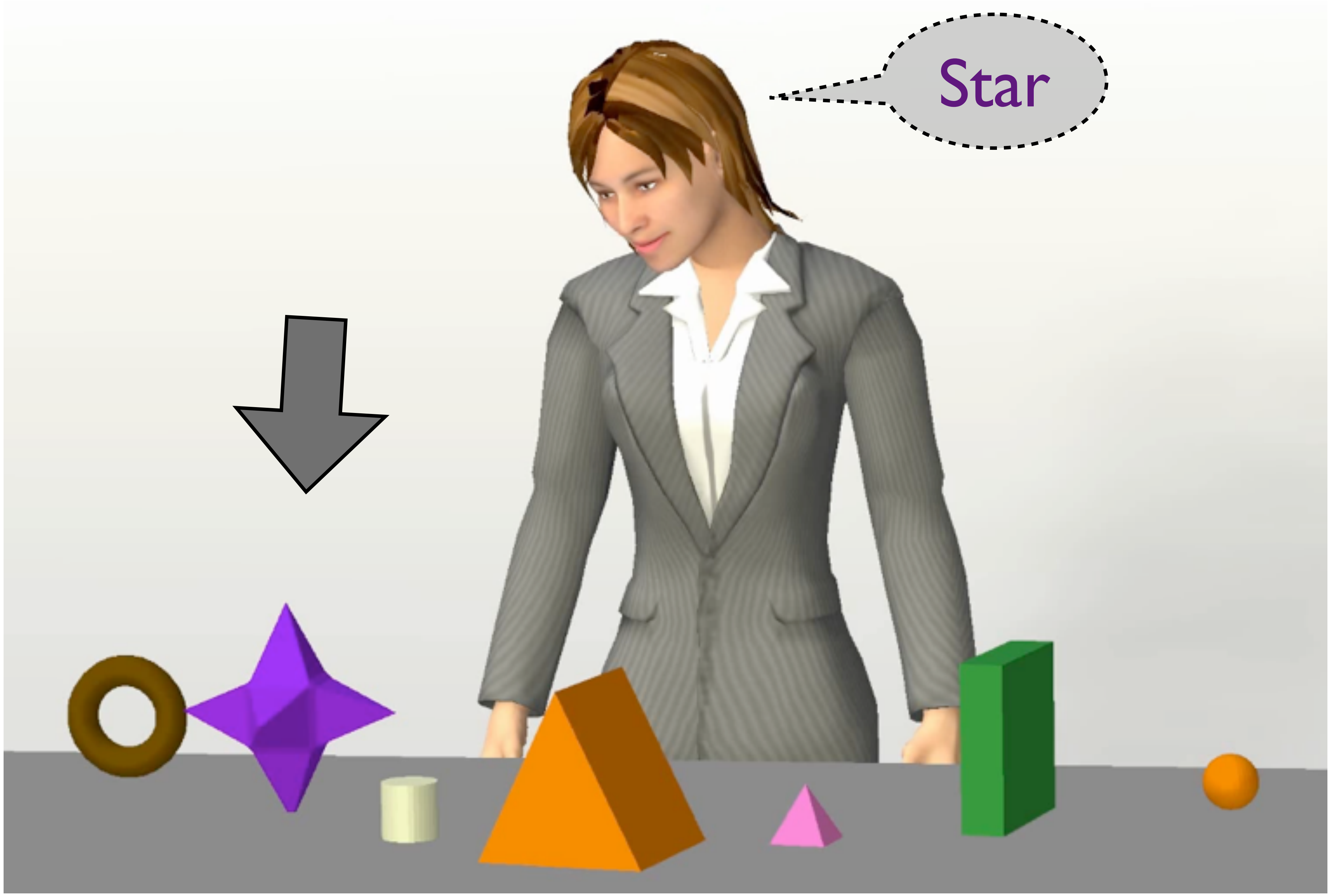


“The star is taller than the pyramid.”

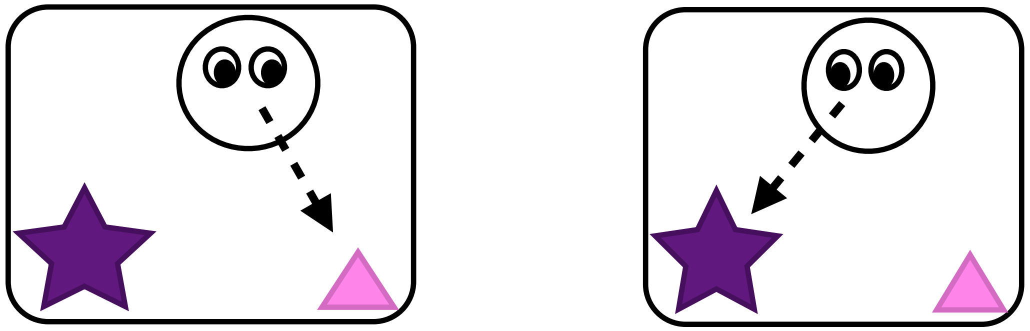
- Reverse cues:
  - Gaze-mediated fixations to “pyramid”
  - Speech-mediated fixations to “star”

➡ **RT data reveals disruption instead of facilitation!**

- What causes the slowed response time?


Star

Experiment 1 : Results



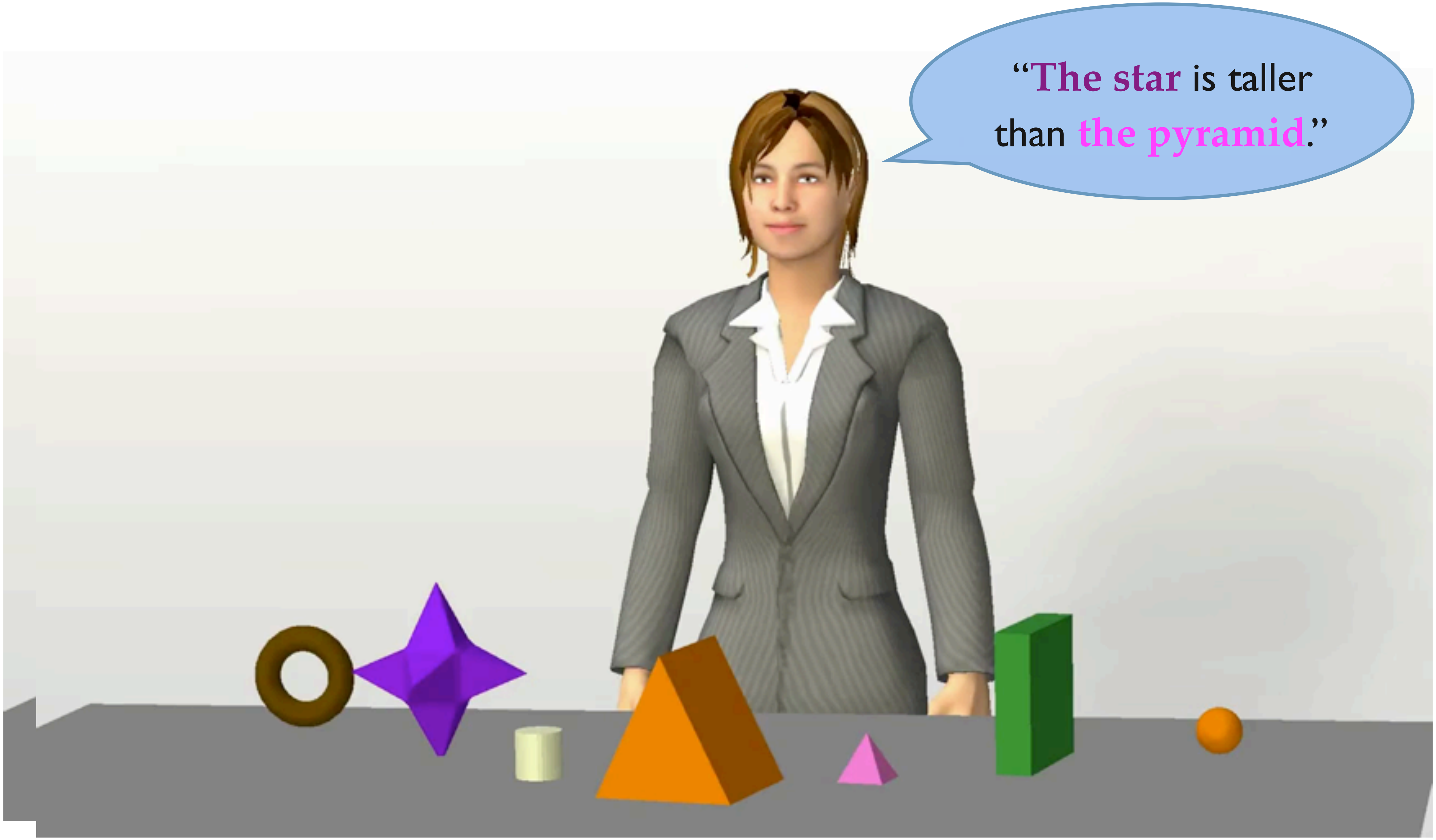
"The star is taller than the pyramid."

- Reverse cues:
  - Gaze-mediated fixations to "pyramid"
  - Speech-mediated fixations to "star"

➡ **RT data reveals disruption instead of facilitation!**

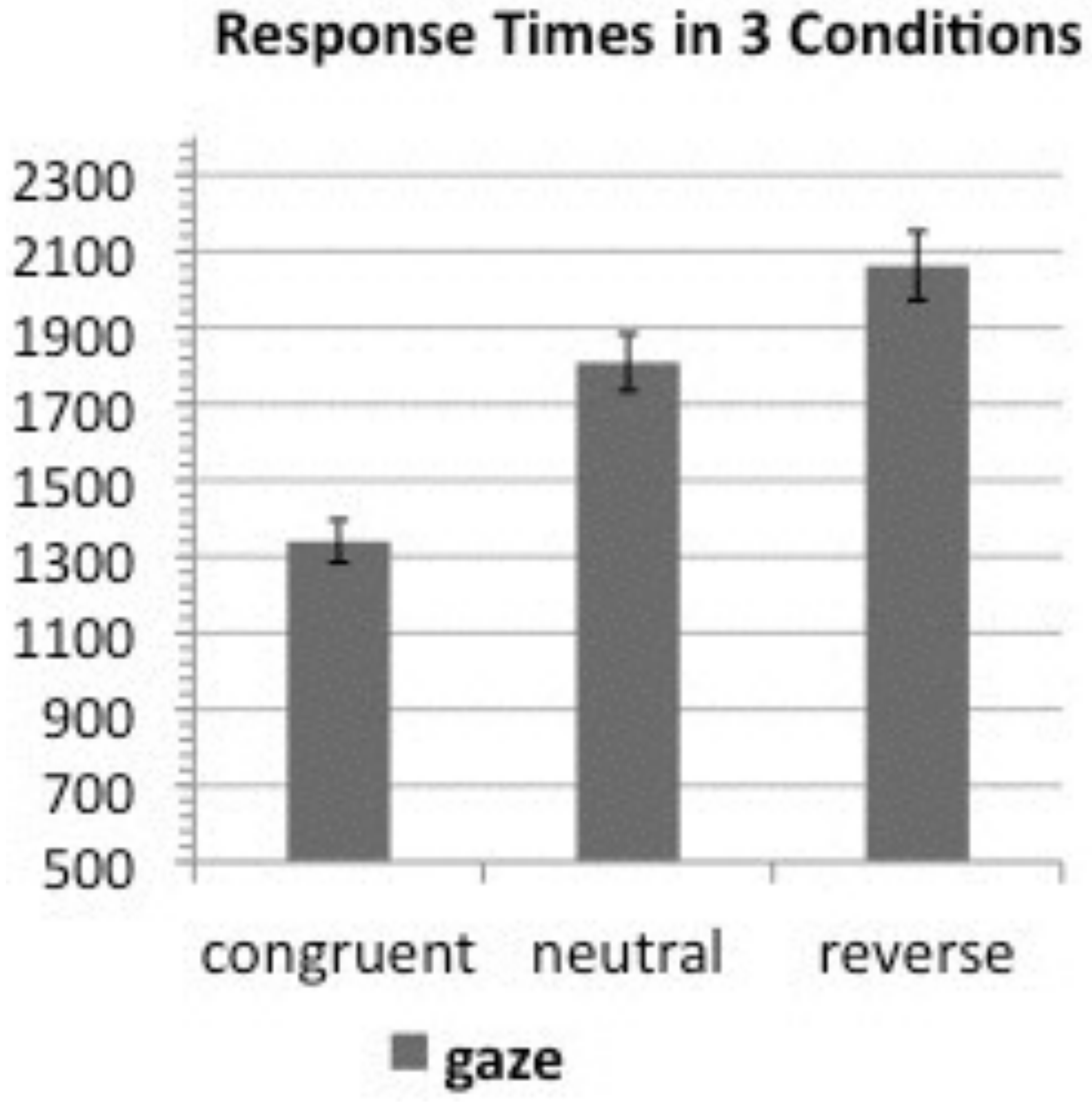
- What causes the slowed response time?
  - Timing and saliency? Or referential intentions?
- Do other (purely) visual cues have the same effect?

Experiment 2

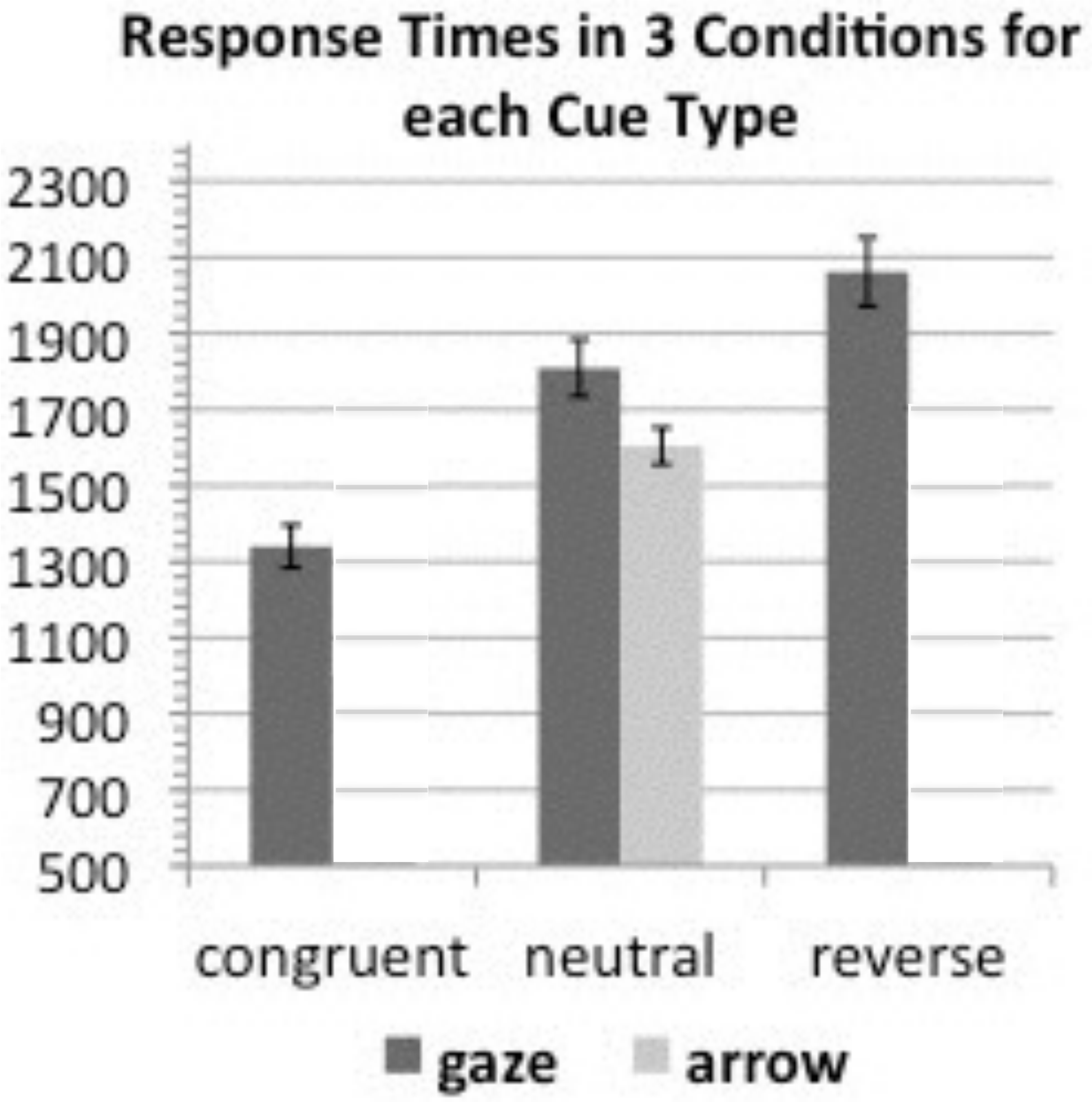


"**The star** is taller than **the pyramid**."

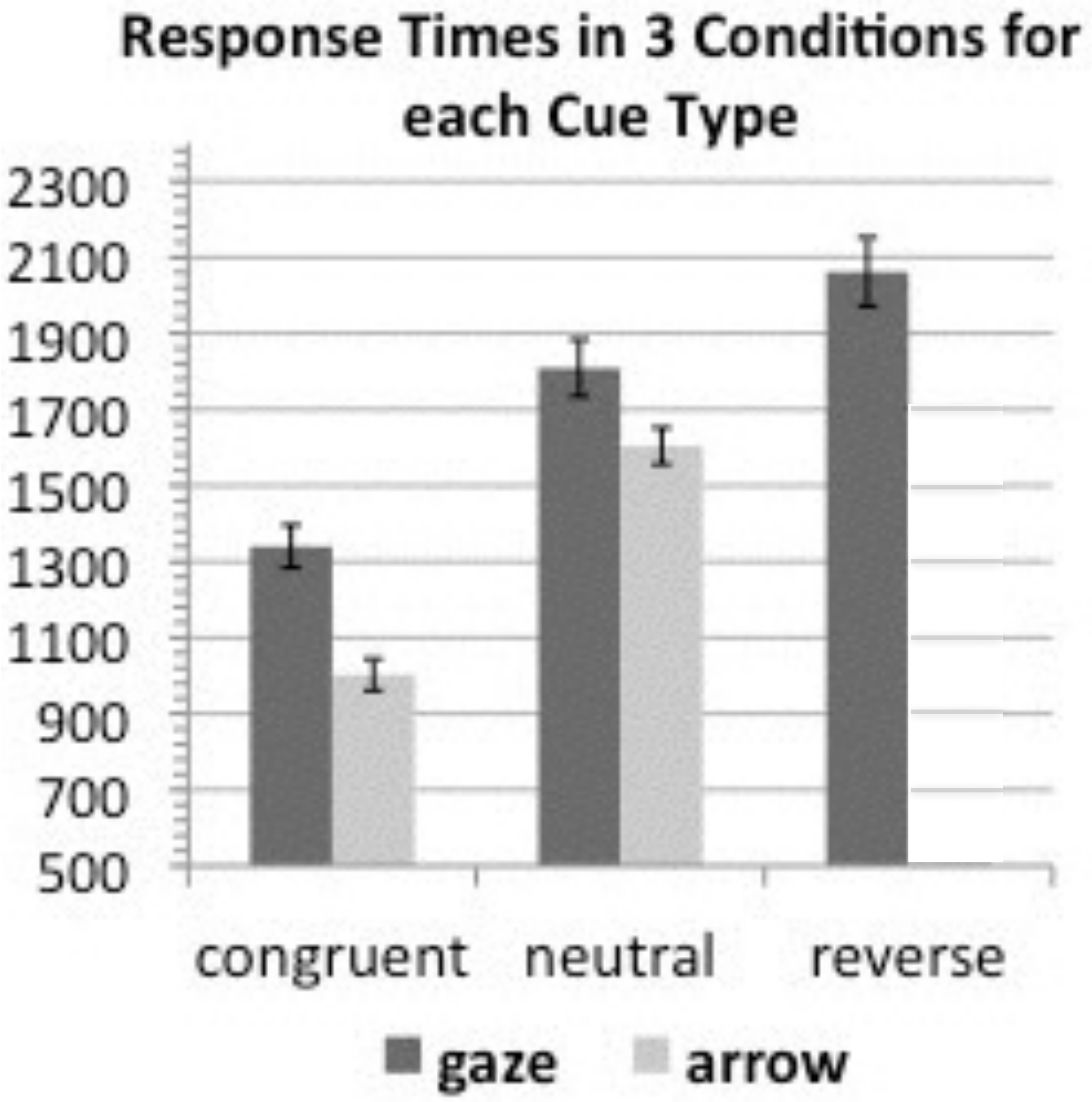
Experiment 1: Results



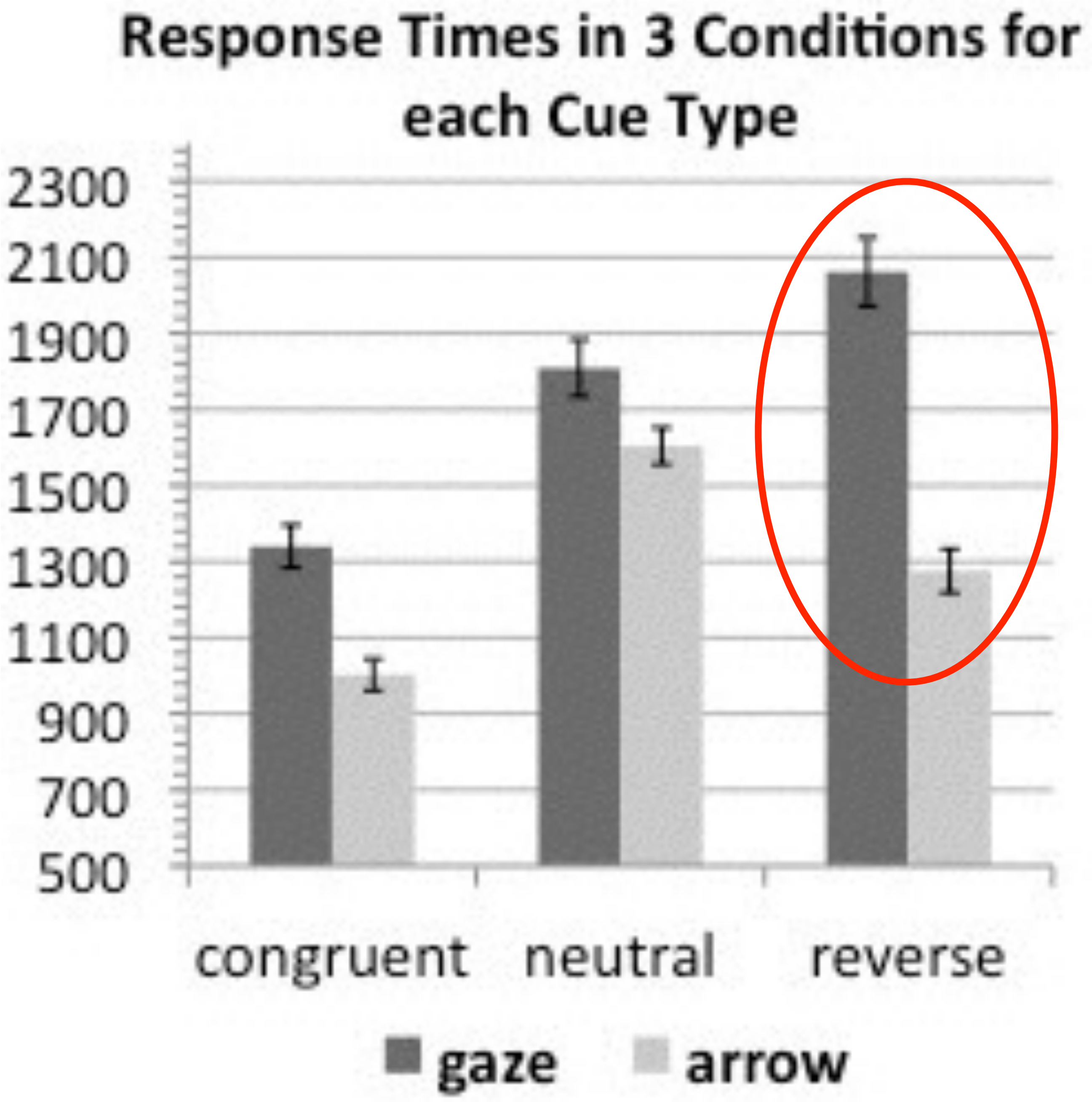
Experiment 1+2 : Results



Experiment 1+2 : Results

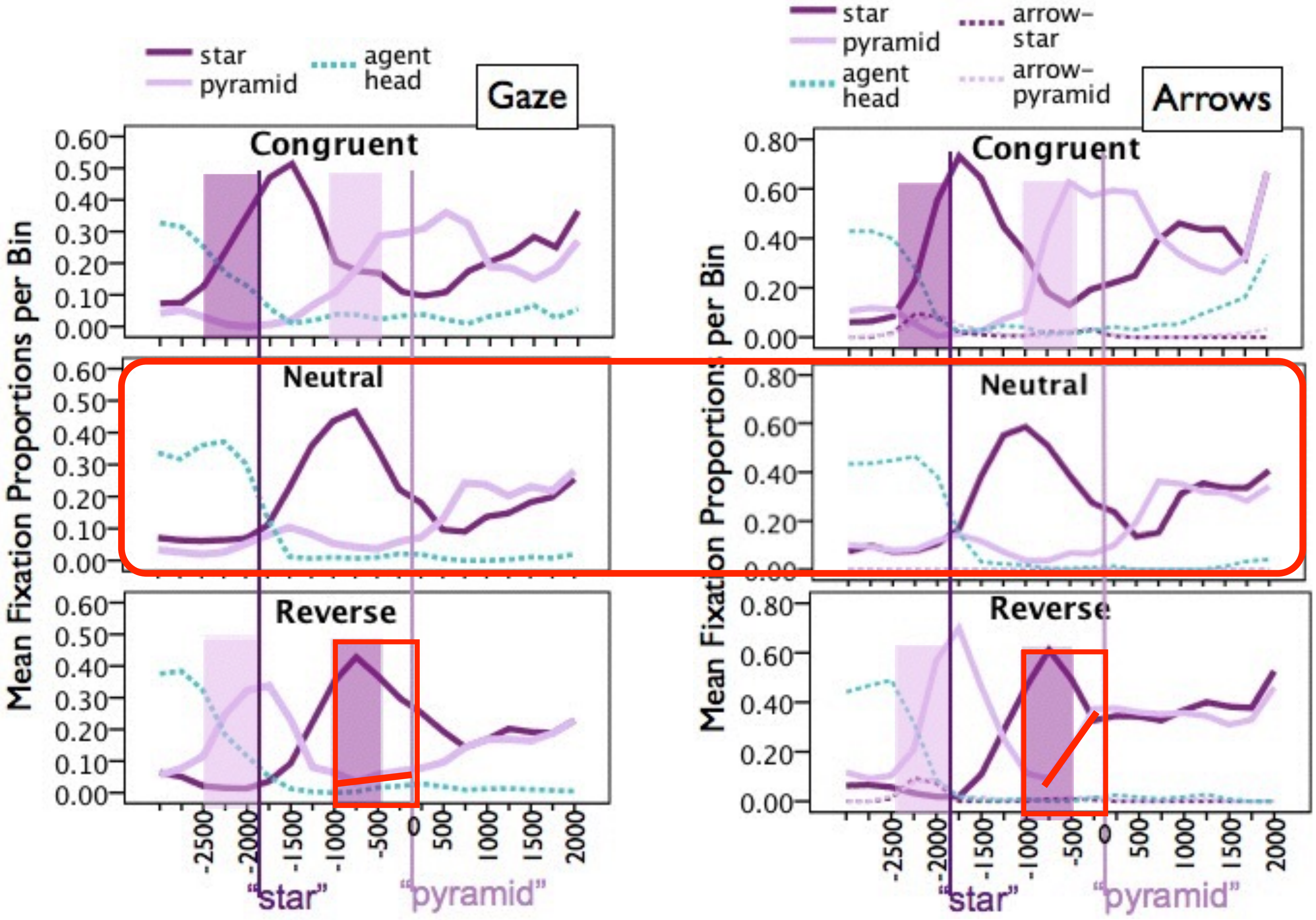


Experiment 1+2 : Results

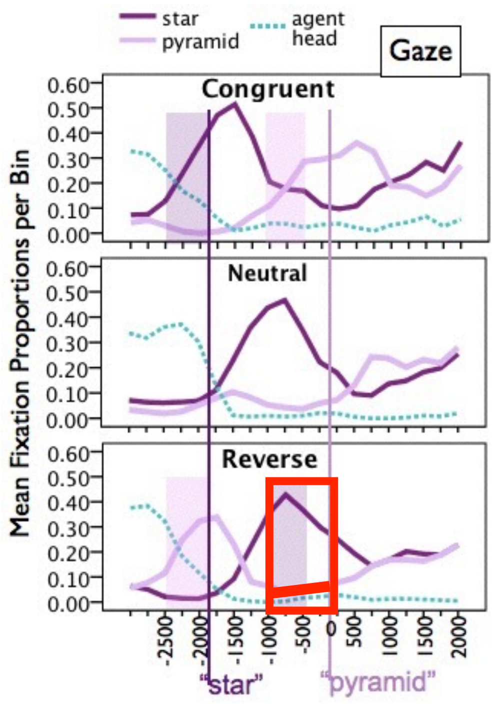


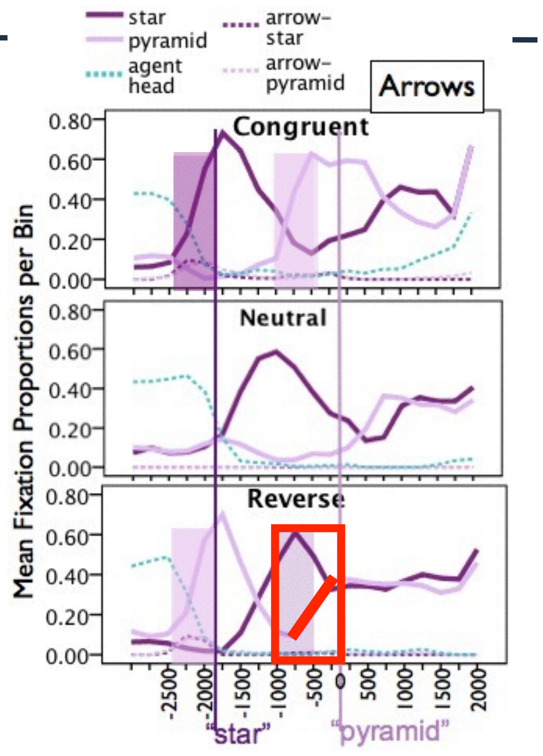
Cue Type x Condition

Experiment 1+2: Results



Further Results





- Response Time Block Analysis:
  - Learning effect for <u>reverse</u> arrows (interaction)
  - No learning effect for <u>reverse</u> gaze (no interaction)

Interim summary

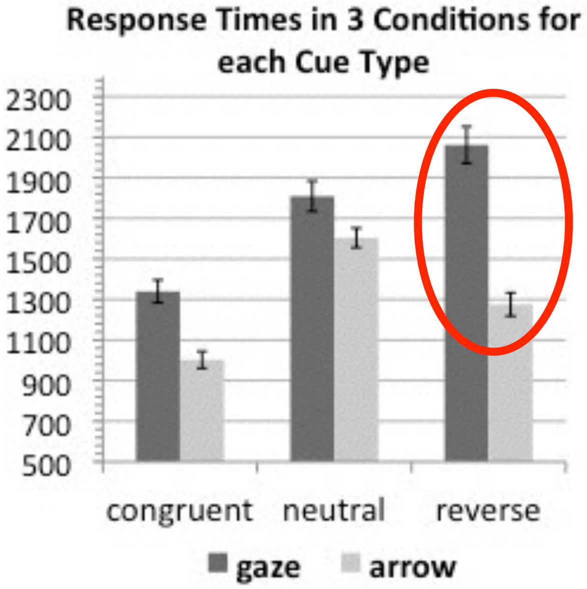
- Gaze elicits a prediction for the next referent
- Strong bias to infer referential intentions as acquired across many years
- Arrows are assigned a task-specific utility
- Unbiased cue which can be used flexibly
- Gaze affects comprehension *beyond* visual cueing

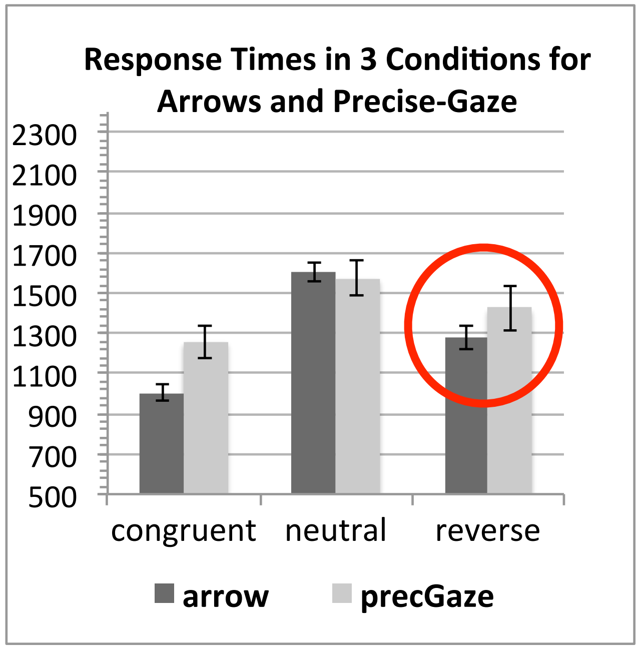
Remaining issues

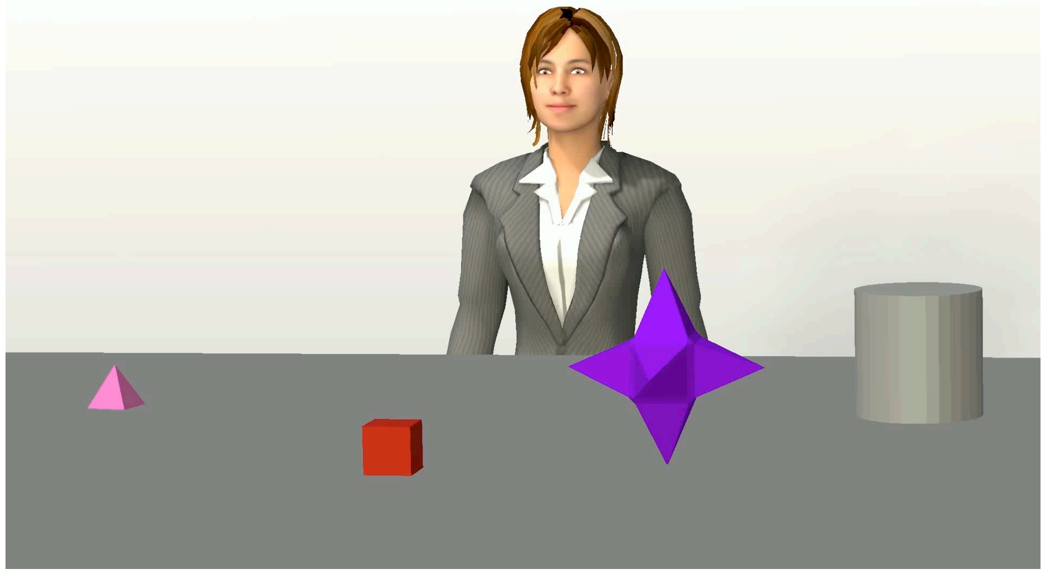
- Difference between speaker gaze and arrow cues:
  - Precision of cue
  - Reliability wrt language

Precision?

- Gaze is (often?) less precise than e.g. arrow cues
- Compare arrows against simplified, precise gaze cue
  - ➡ Benefit in “reverse” condition!



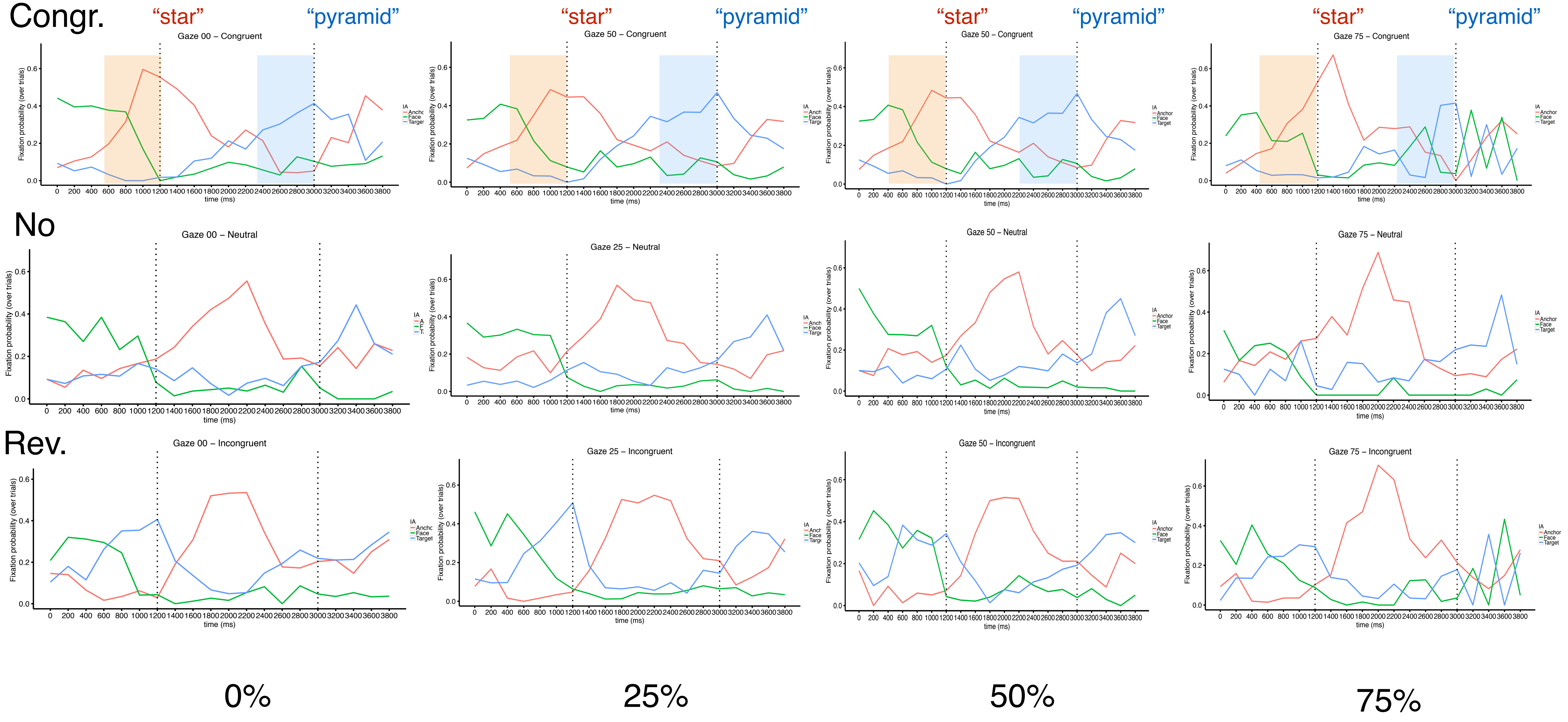




Reliability

- Gaze occurs more often / more naturally than arrows
- Tendency to trust & follow gaze more than arrows?
    - Arrow usage more strategic?
    - Compare 0% (as before) with 25%, 50%, 75% trials with **invalid** cues in experiment

Reliability - Cue following



Congr.

“star”

“pyramid”

No

Rev.

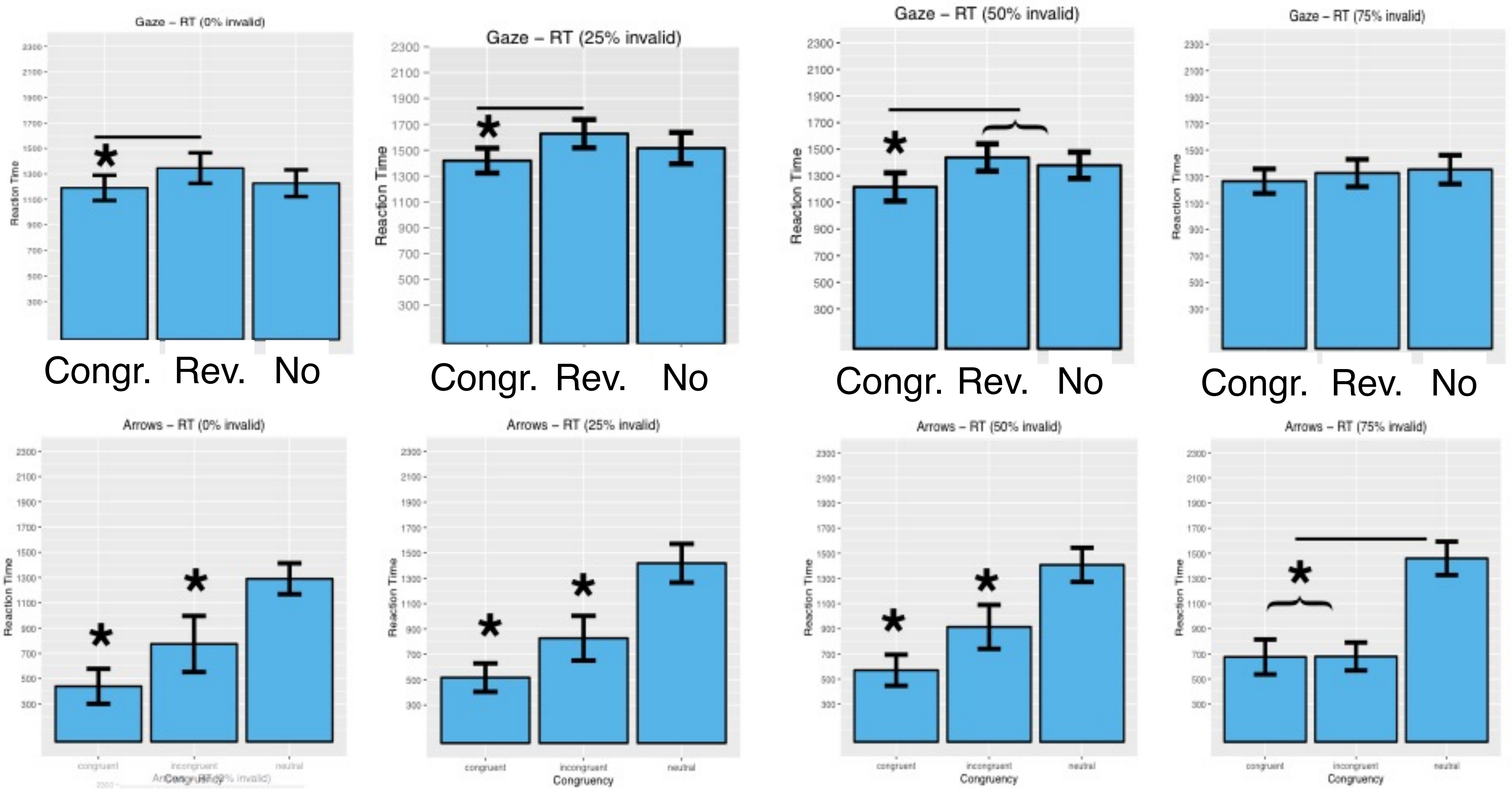
0%

25%

50%

75%

Reliability - Cue effect



Congr. Rev. No

Congr. Rev. No

Congr. Rev. No

Congr. Rev. No

0%

25%

50%

75%

Reliability

- Listeners stop following gaze (only) when cue is misleading in 75% of trials
- Listeners keep following and benefiting from arrows (even when cue misleads in 75%) !

➡ Gaze-following is less automatic

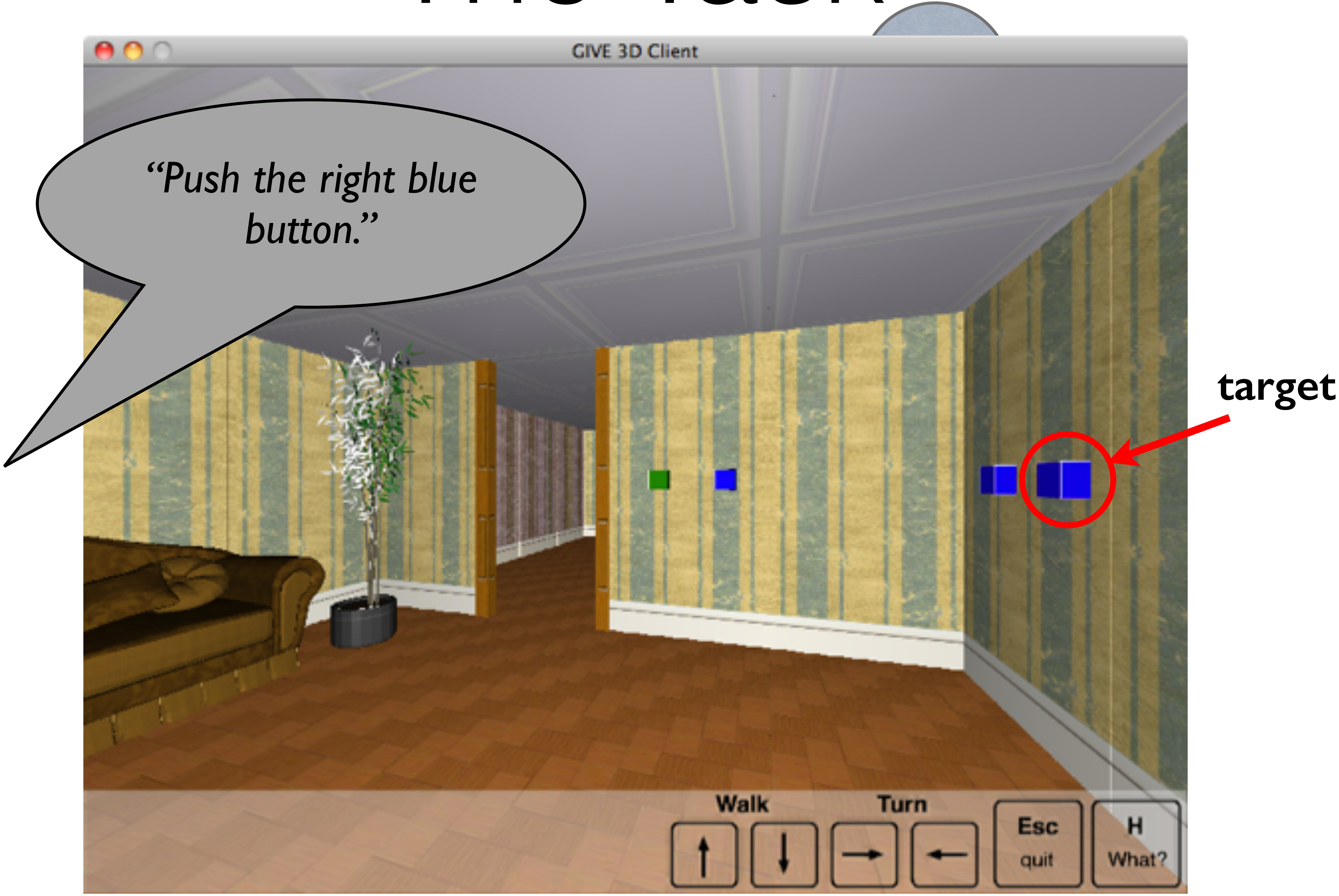
➡ But also less strategic than arrow usage

Listener gaze

Listener Gaze

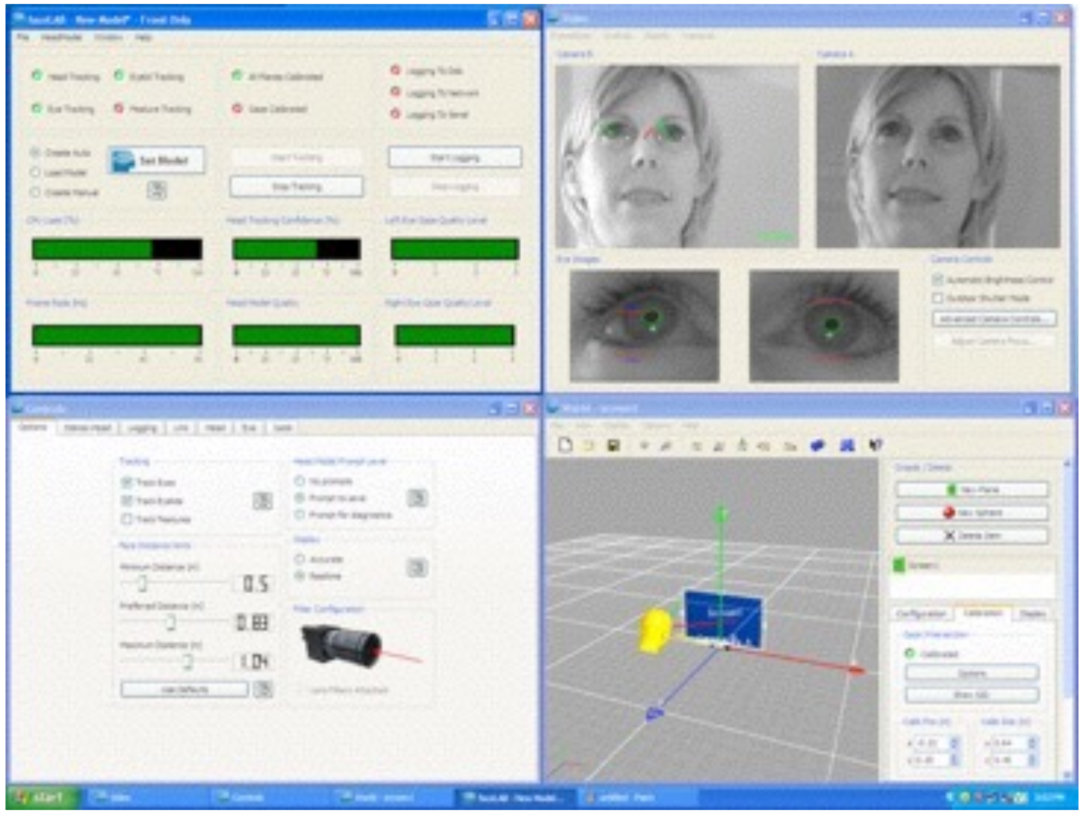
- Listeners look at
  - what they hear
  - what the speaker looks at
- Speakers monitor what listeners look at
- How can/do they exploit this information?
  - Can we evaluate instructions (better) using eye-tracking?
  - Can we construct instructions (better) using eye-tracking?

The Task



GIVE setting (Koller et al., 2010)

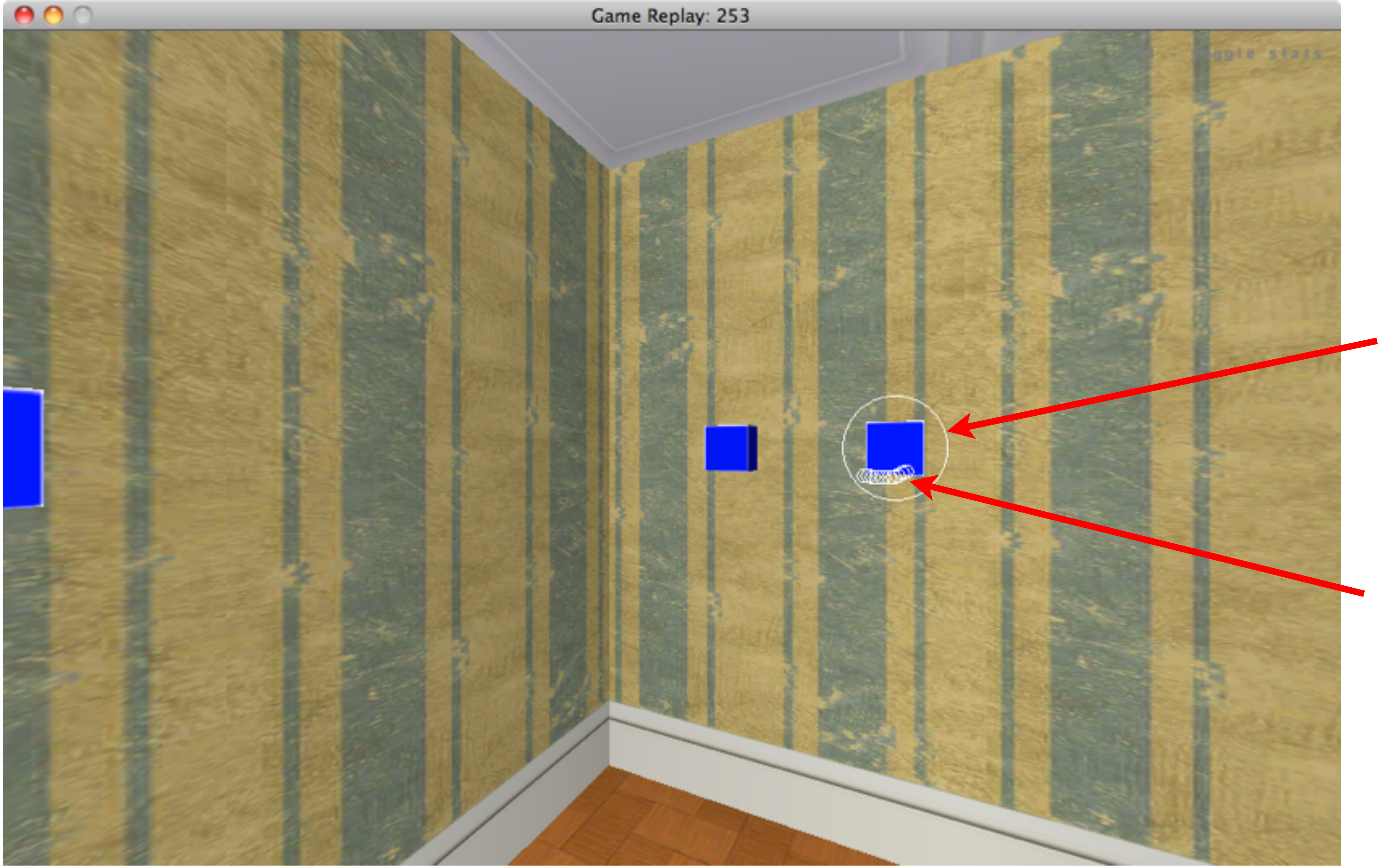
Recording object inspections



faceLAB eye-tracking system

- Every 15ms, sample 2D position on screen that the user is fixating
- Resolve this position to the corresponding object in the current 3D scene
- User looks to an object of more than 300ms count as a “referential inspection” of that object

Tracking listener gaze



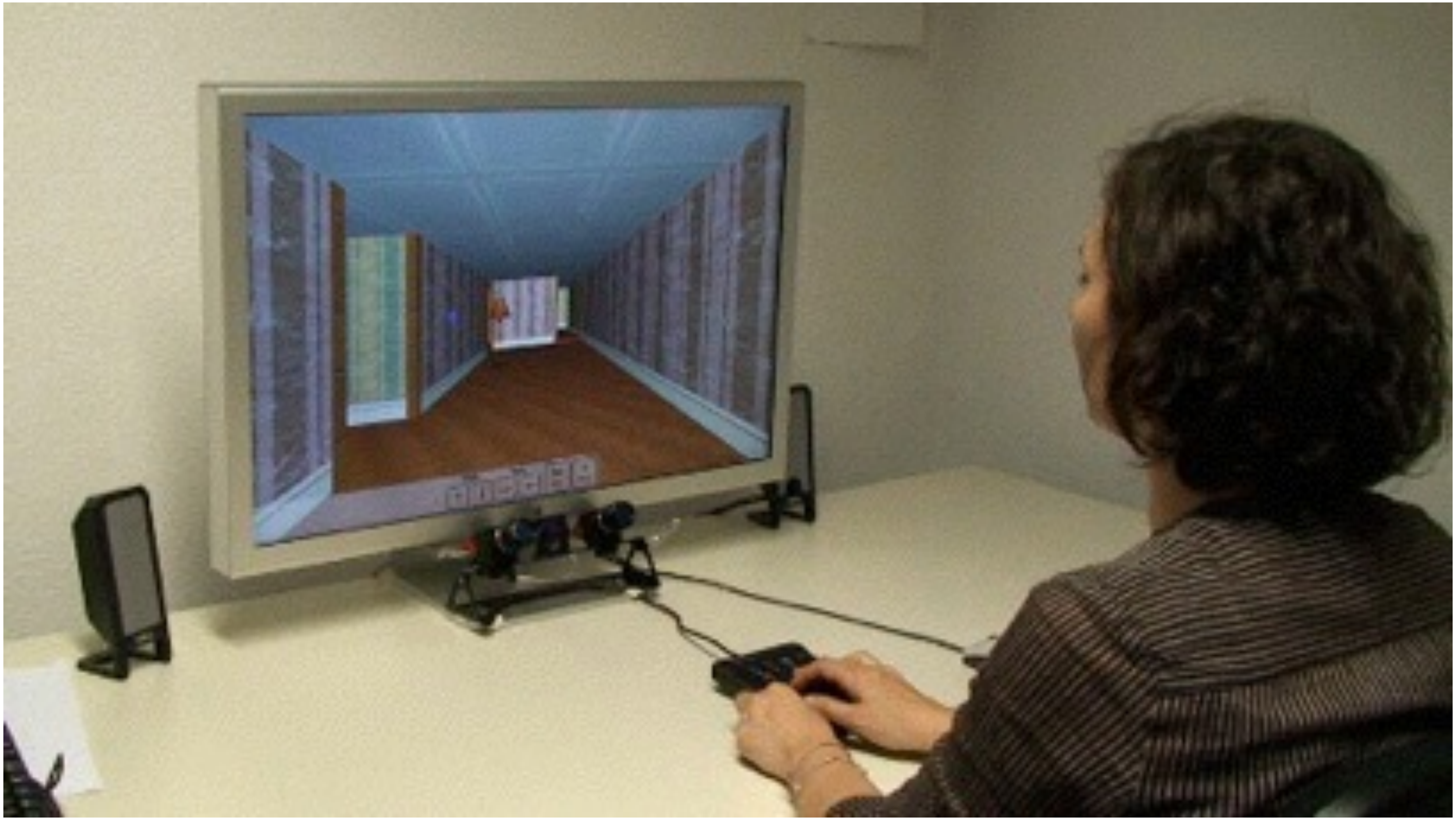
inspection of object

trace of fixations

Monitoring understanding

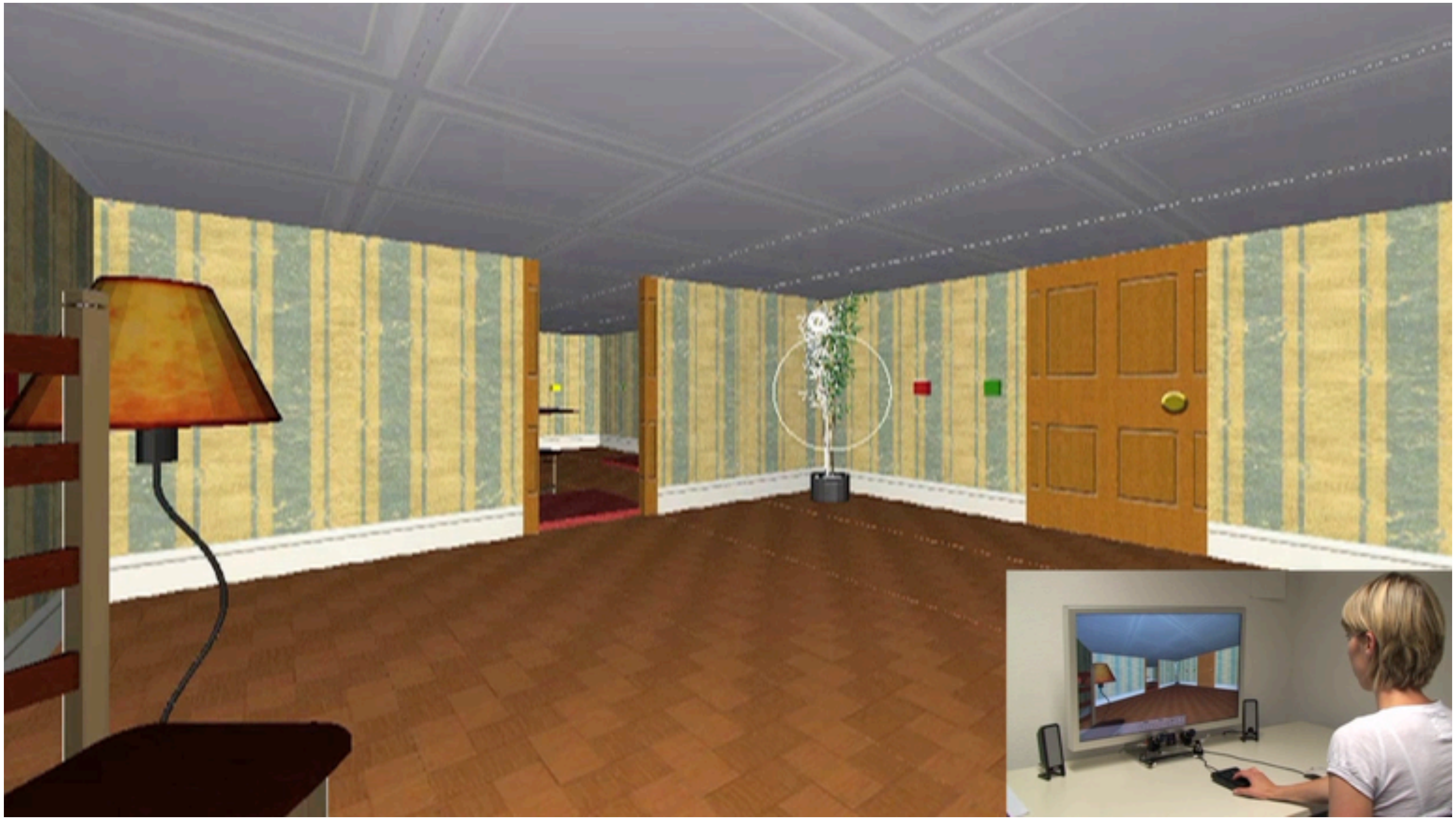
- Based on eye gaze, system attempts to predict whether the user has understood its referring expressions
- System generates proactive feedback accordingly
    - Target inspection: “Yes, that one!”
    - Distractor inspection: “No, not that one!”

Setup



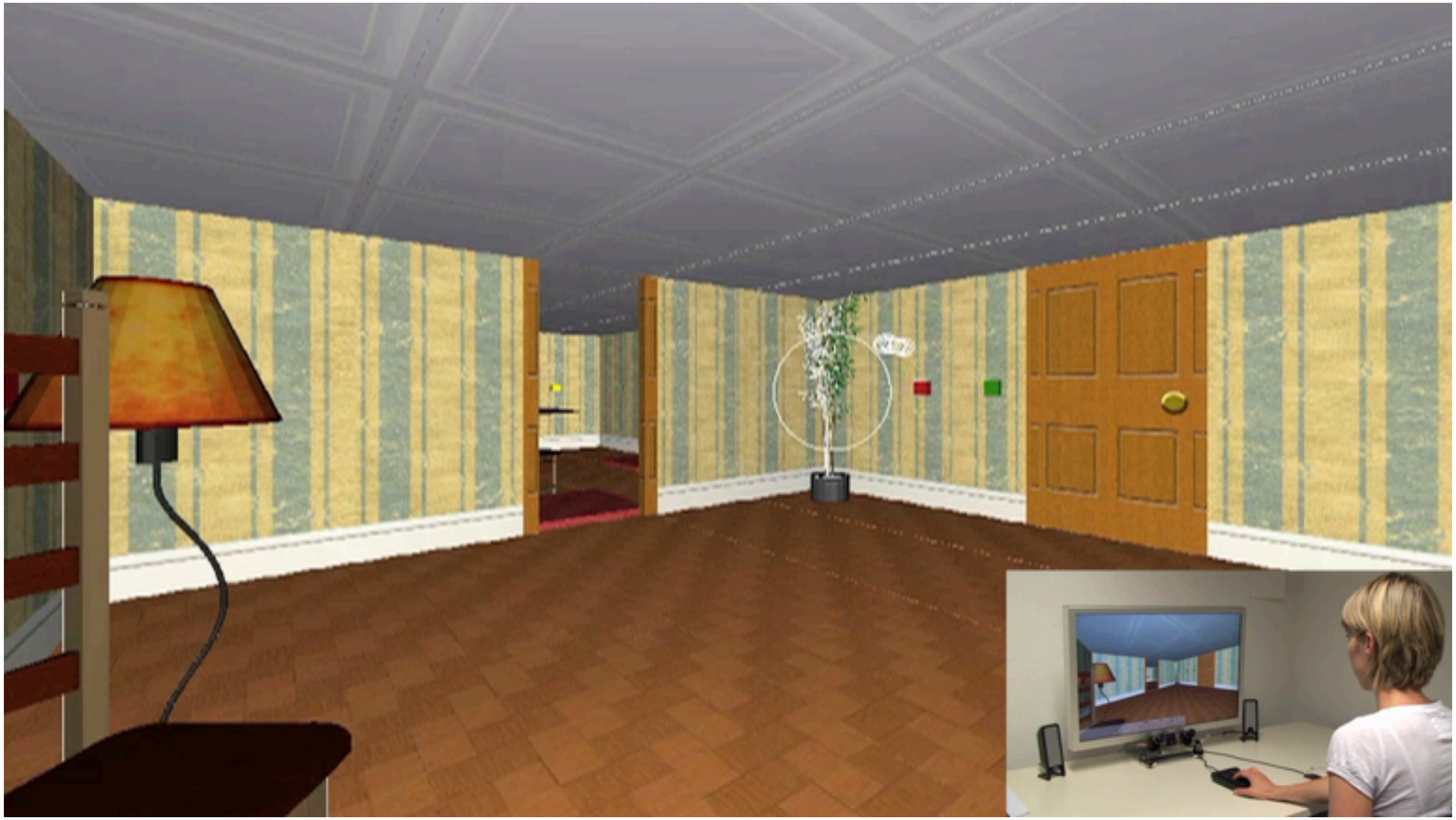
Example scene

“Push the left button to the...“



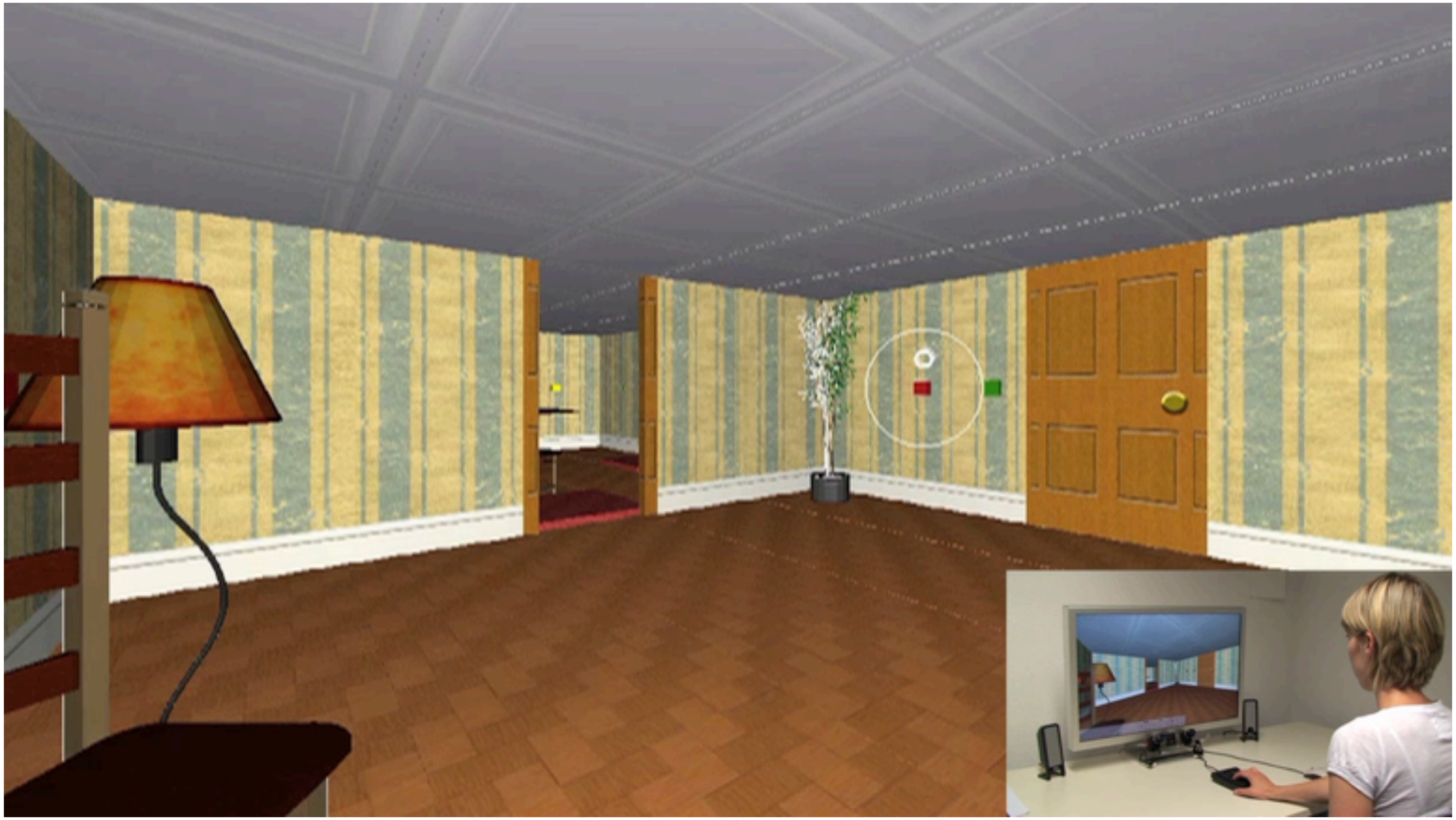
Example scene

“...right of the flower.   “



Example scene

“...flower.   - Yes, that one.“



Baseline 1:
No feedback

- No monitoring of referential understanding
- No proactive feedback
- System generates a follow-up referring expression only after user has pressed wrong button or asked for help

Baseline 2:
Movement-based feedback

- System makes prediction only if user moves towards single visible button
- Same feedback as gaze-based system
    - Movement towards target: “Yes, that one!”
    - Movement towards distractor: “No, not that one!”

Eye movements

	inspection durations	
	target	distractor
successful		
eyetracking	2111.6	720.5
no-feedback	1492.0***	185.7***
movement	1493.8**	260.5***
unsuccessful		
eyetracking	752.1	3378.9
no-feedback	619.5	1891.7
movement	602.6	2113.1

*Differences to eyetracking system statistically significant at ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$*

Eye movements

	inspection durations	
	target	distractor
successful		
eyetracking	**2111.6**	720.5
no-feedback	**1492.0*** **	185.7***
movement	**1493.8****	260.5***
unsuccessful		
eyetracking	752.1	**3378.9**
no-feedback	619.5	**1891.7**
movement	602.6	**2113.1**

listeners spend more time looking at what they think is the referent than at other buttons

Interaction Effectiveness

average number of help requests per interaction

	confusion	success
eyetracking	1.14	91.9
no-feedback	2.26**	83.5**
movement	1.77*	87.5

gaze-based feedback makes users more confident in the interaction

Interaction Effectiveness

proportion of correctly resolved referring expressions

	confusion	success
eyetracking	1.14	91.9
no-feedback	2.26**	83.5**
movement	1.77*	87.5

tracking listener's gaze enhances referential success

Interim summary II

- Listeners reliably inspect understood referents in all conditions
- Gaze feedback results in:
    - Lower confusion
    - Positive feedback: speeds interaction
    - Negative feedback: increases success
- (But timing remains an issue!)

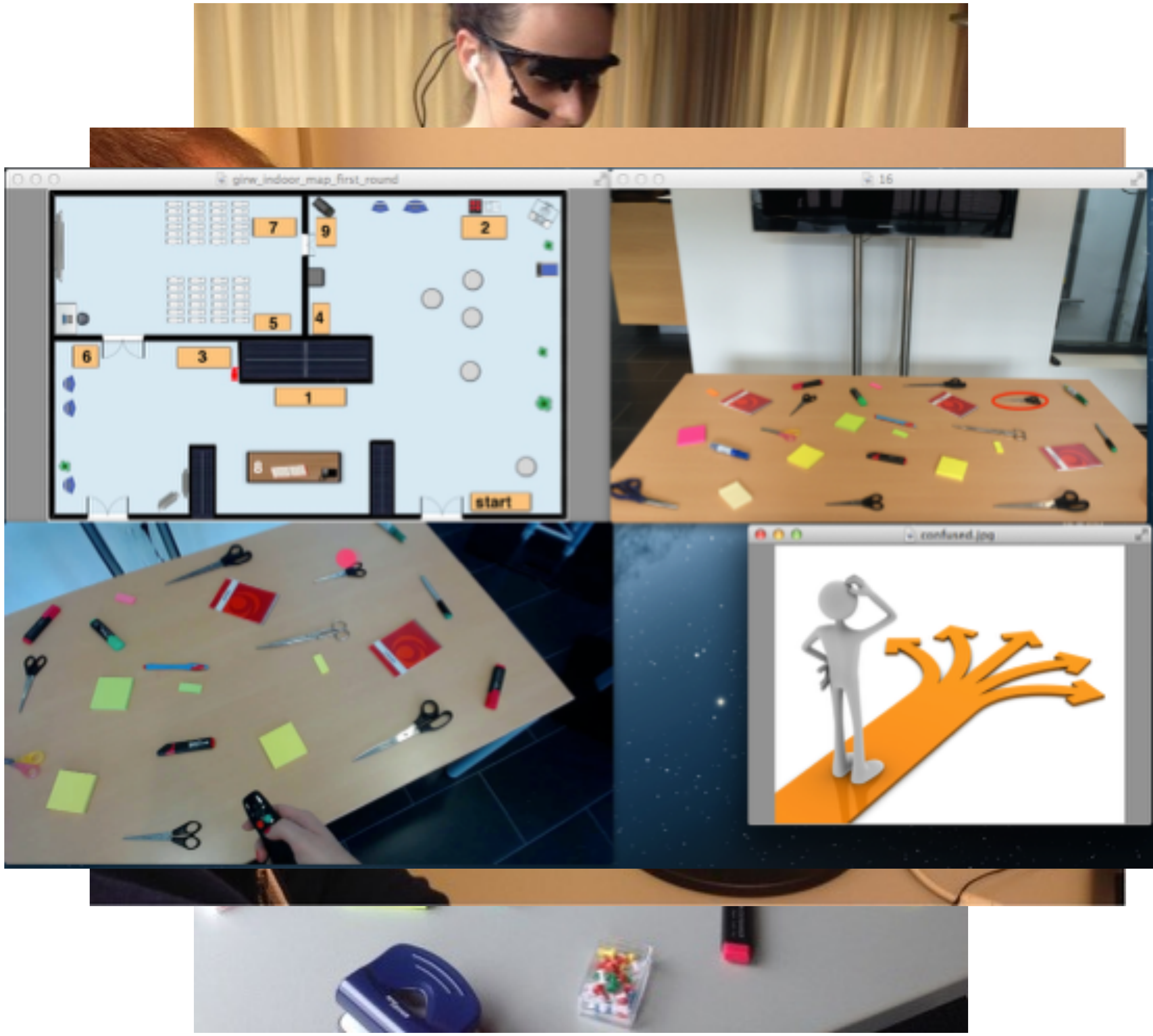
Timing

Human speaker?

- Is this how human speakers use listener gaze?
- Which eye-movements do they rely on?
- What does feedback really look like?

Setup

- Walker (12 pairs)
  - Unknown location
  - Eye-tracked by PUPIL P
  - Hears instructions
- Instructor
  - Map
  - Sees walker scene view
  - Gives instructions



Task

- Walker needs to find table (makro)
- Then walker takes objects and puts them aside (mikro)
- 9 thematic tables with 3-4 target objects each
  - 3 tables in each condition
  - ~40min total

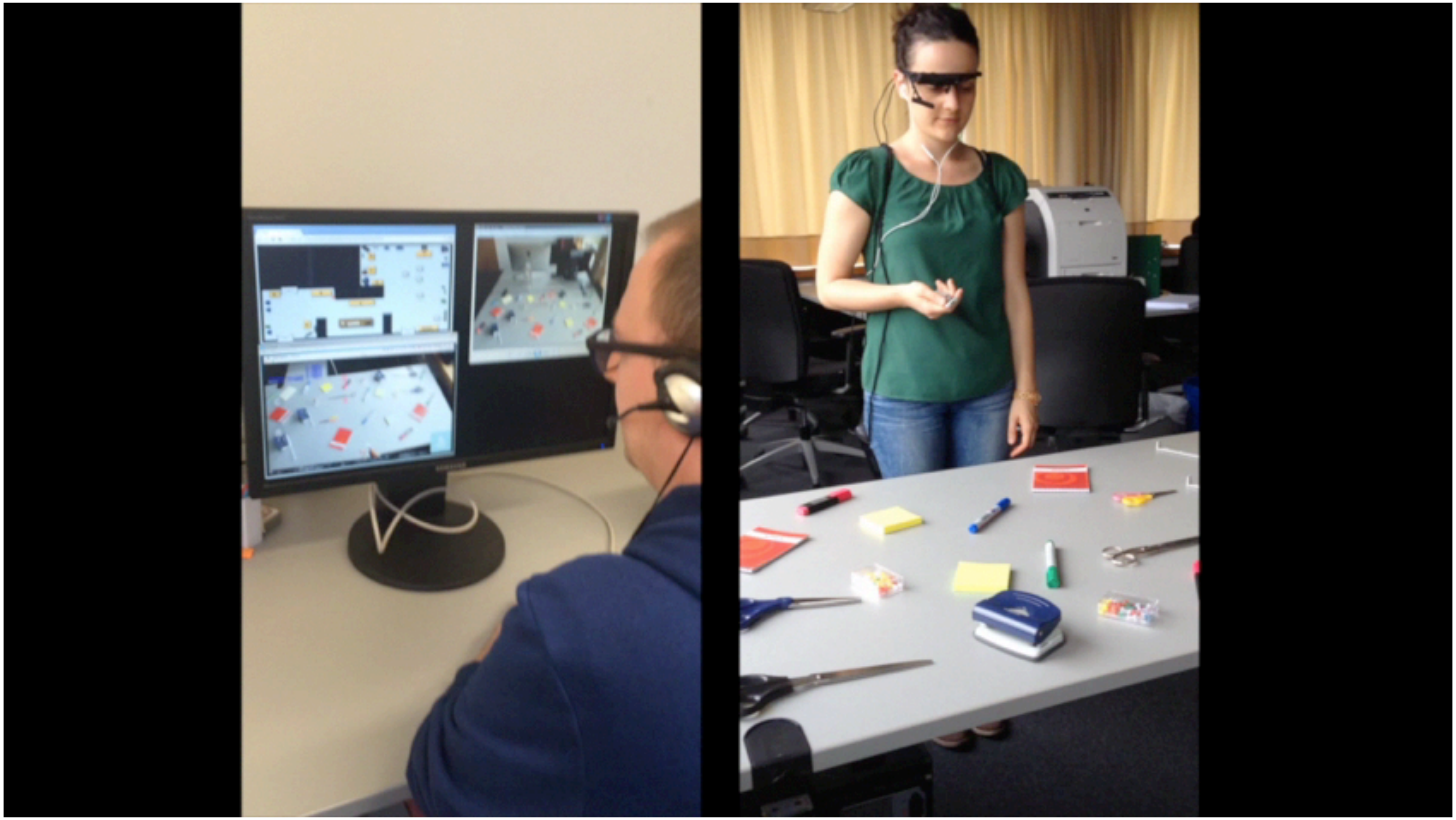
Conditions

1.GAZE : Walker gaze available to instructor

2.Man-GAZE : Walker gaze perturbed (20% random shift)

3.No-GAZE: Walker gaze NOT available

Setup



“Could you please pick up the pin box .. eh… that’s furthest away from you.”

Measures

Dependent Variables (DV)

1. Instructor behavior:

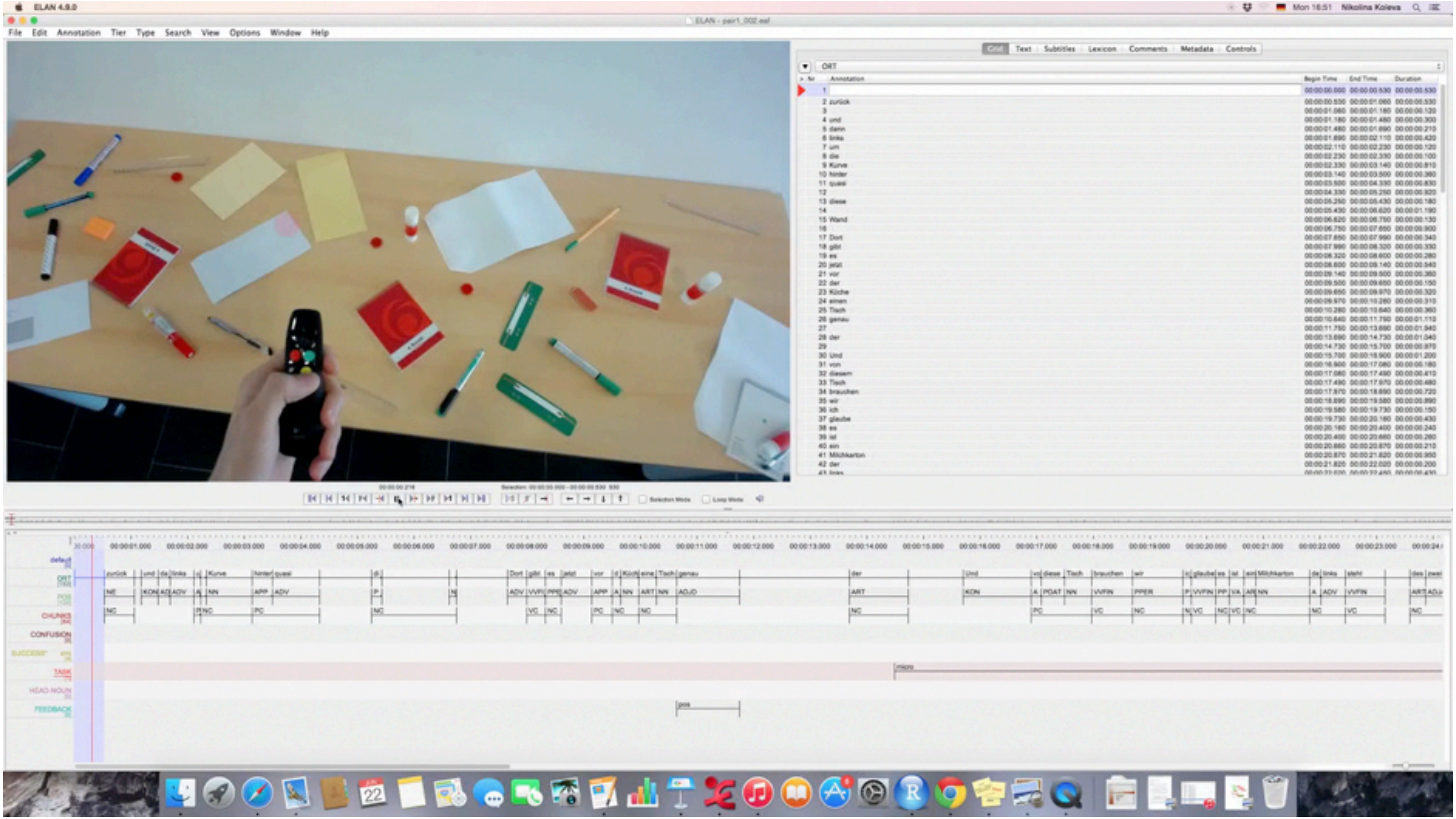
   a) no. words, feedback (Q1)

2. Low-level listener eye-movements (Q2-i)

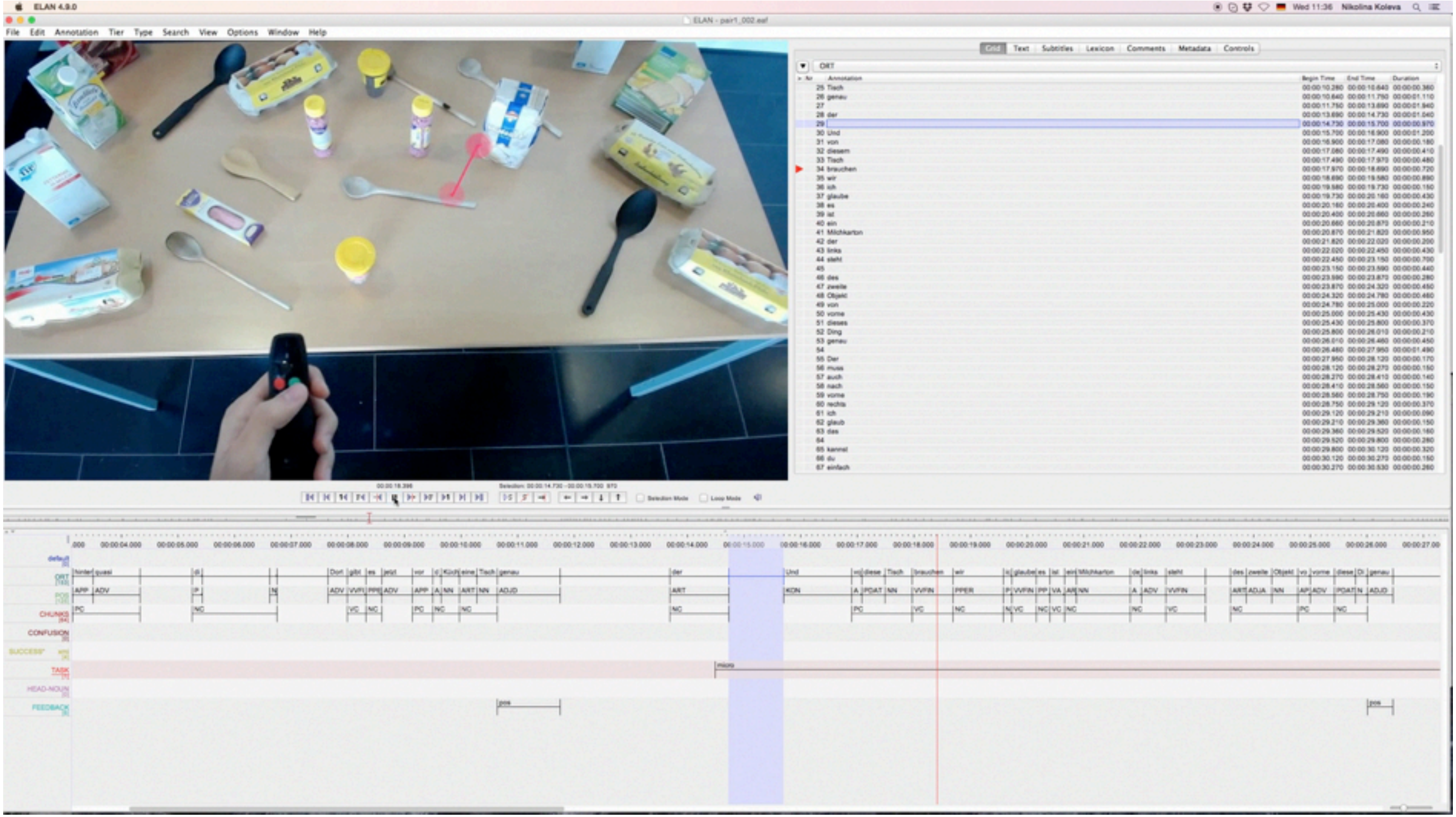
3. Listener eye-movements in relation to feedback (Q2-ii)

Preprocessing DV1

- Transcription
- Forced alignment
- Automatic lemmatization, tagging, shallow syntactic analyses (TreeTagger, Schmid 1995)
- Semi-manual annotation of feedback instances (neg. & pos.)



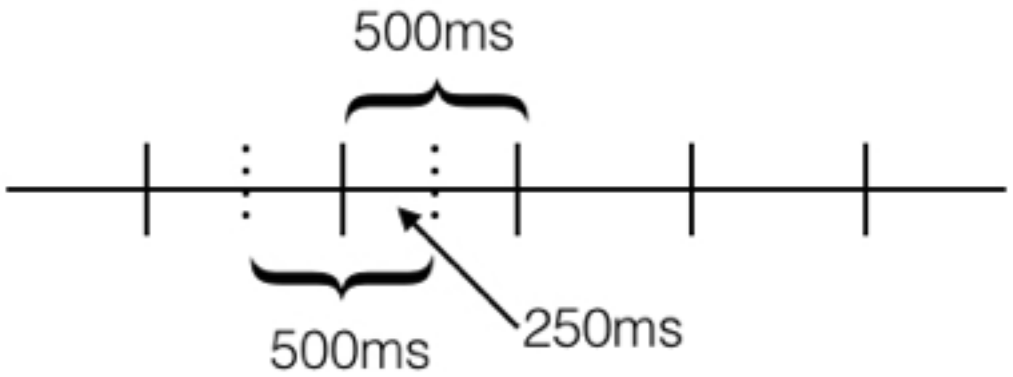
Nr	Annotation	Begin Time	End Time	Duration
1		00:00:00.000	00:00:00.530	00:00:00.530
2	zurück	00:00:00.530	00:00:01.060	00:00:00.530
3		00:00:01.060	00:00:01.180	00:00:00.120
4	und	00:00:01.180	00:00:01.480	00:00:00.300
5	dann	00:00:01.480	00:00:01.690	00:00:00.210
6	links	00:00:01.690	00:00:02.110	00:00:00.420
7	um	00:00:02.110	00:00:02.230	00:00:00.120
8	die	00:00:02.230	00:00:02.330	00:00:00.100
9	Kurve	00:00:02.330	00:00:03.140	00:00:00.810
10	hinter	00:00:03.140	00:00:03.500	00:00:00.360
11	quasi	00:00:03.500	00:00:04.330	00:00:00.830
12		00:00:04.330	00:00:05.250	00:00:00.920
13	diese	00:00:05.250	00:00:05.430	00:00:00.180
14		00:00:05.430	00:00:06.620	00:00:01.190
15	Wand	00:00:06.620	00:00:06.750	00:00:00.130
16		00:00:06.750	00:00:07.650	00:00:00.900
17	Dort	00:00:07.650	00:00:07.990	00:00:00.340
18	gibt	00:00:07.990	00:00:08.320	00:00:00.330
19	es	00:00:08.320	00:00:08.600	00:00:00.280
20	jetzt	00:00:08.600	00:00:09.140	00:00:00.540
21	vor	00:00:09.140	00:00:09.500	00:00:00.360
22	der	00:00:09.500	00:00:09.650	00:00:00.150
23	Küche	00:00:09.650	00:00:09.970	00:00:00.320
24	einen	00:00:09.970	00:00:10.280	00:00:00.310
25	Tisch	00:00:10.280	00:00:10.640	00:00:00.360
26	genau	00:00:10.640	00:00:11.750	00:00:01.110
27		00:00:11.750	00:00:13.690	00:00:01.940
28	der	00:00:13.690	00:00:14.730	00:00:01.040
29		00:00:14.730	00:00:15.700	00:00:00.970
30	Und	00:00:15.700	00:00:16.900	00:00:01.200
31	von	00:00:16.900	00:00:17.080	00:00:00.180
32	diesem	00:00:17.080	00:00:17.490	00:00:00.410
33	Tisch	00:00:17.490	00:00:17.970	00:00:00.480
34	brauchen	00:00:17.970	00:00:18.690	00:00:00.720
35	wir	00:00:18.690	00:00:19.580	00:00:00.890
36	ich	00:00:19.580	00:00:19.730	00:00:00.150
37	glaube	00:00:19.730	00:00:20.160	00:00:00.430
38	es	00:00:20.160	00:00:20.400	00:00:00.240
39	ist	00:00:20.400	00:00:20.660	00:00:00.260
40	ein	00:00:20.660	00:00:20.870	00:00:00.210
41	Milchkarton	00:00:20.870	00:00:21.820	00:00:00.950
42	der	00:00:21.820	00:00:22.020	00:00:00.200
43	links	00:00:22.020	00:00:22.450	00:00:00.430



Nr	Annotation	Begin Time	End Time	Duration
25	Tisch	00:00:10.280	00:00:10.640	00:00:00.360
26	genau	00:00:10.640	00:00:11.750	00:00:01.110
27		00:00:11.750	00:00:13.690	00:00:01.940
28	der	00:00:13.690	00:00:14.730	00:00:01.040
29		00:00:14.730	00:00:15.700	00:00:00.970
30	Und	00:00:15.700	00:00:16.900	00:00:01.200
31	von	00:00:16.900	00:00:17.080	00:00:00.180
32	diesem	00:00:17.080	00:00:17.490	00:00:00.410
33	Tisch	00:00:17.490	00:00:17.970	00:00:00.480
34	brauchen	00:00:17.970	00:00:18.690	00:00:00.720
35	wir	00:00:18.690	00:00:19.580	00:00:00.890
36	ich	00:00:19.580	00:00:19.730	00:00:00.150
37	glaube	00:00:19.730	00:00:20.160	00:00:00.430
38	es	00:00:20.160	00:00:20.400	00:00:00.240
39	ist	00:00:20.400	00:00:20.660	00:00:00.260
40	ein	00:00:20.660	00:00:20.870	00:00:00.210
41	Milchkarton	00:00:20.870	00:00:21.820	00:00:00.950
42	der	00:00:21.820	00:00:22.020	00:00:00.200
43	links	00:00:22.020	00:00:22.450	00:00:00.430
44	steht	00:00:22.450	00:00:23.150	00:00:00.700
45		00:00:23.150	00:00:23.590	00:00:00.440
46	des	00:00:23.590	00:00:23.870	00:00:00.280
47	zweite	00:00:23.870	00:00:24.320	00:00:00.450
48	Objekt	00:00:24.320	00:00:24.780	00:00:00.460
49	von	00:00:24.780	00:00:25.000	00:00:00.220
50	vorne	00:00:25.000	00:00:25.430	00:00:00.430
51	dieses	00:00:25.430	00:00:25.800	00:00:00.370
52	Ding	00:00:25.800	00:00:26.010	00:00:00.210
53	genau	00:00:26.010	00:00:26.460	00:00:00.450
54		00:00:26.460	00:00:27.950	00:00:01.490
55	Der	00:00:27.950	00:00:28.120	00:00:00.170
56	muss	00:00:28.120	00:00:28.270	00:00:00.150
57	auch	00:00:28.270	00:00:28.410	00:00:00.140
58	nach	00:00:28.410	00:00:28.560	00:00:00.150
59	vorne	00:00:28.560	00:00:28.750	00:00:00.190
60	rechts	00:00:28.750	00:00:29.120	00:00:00.370
61	ich	00:00:29.120	00:00:29.210	00:00:00.090
62	glaub	00:00:29.210	00:00:29.360	00:00:00.150
63	das	00:00:29.360	00:00:29.520	00:00:00.160
64		00:00:29.520	00:00:29.800	00:00:00.280
65	kannst	00:00:29.800	00:00:30.120	00:00:00.320
66	du	00:00:30.120	00:00:30.270	00:00:00.150
67	einfach	00:00:30.270	00:00:30.530	00:00:00.260

Preprocessing DV2

- Standard dispersion-based fixation detection algorithm (Salvucci & Goldberg, 2000)
  - “sequence of gaze points to be a fixation if the maximum distance from their joint center is less than 5% of the scene camera width and the sequence has a minimum duration of 66 msec“
- Sliding window (500ms, step size 250ms) to extract eye movement features, resulting in a dataset of 18,841 time windows



Preprocessing DV2

Fixation	rate, mean, max, variance of durations mean, variance of variance within one fix.
Saccades	rate, ratio of (small/large/right/left) sacc. mean, max, variance of amplitudes
Combined	ratio saccades / fixations
Wordbooks	number of non-zero entries maximum and minimum entries as well as their difference for n-grams with $n \leq 4$
Ratios	all fixation, saccade and combined features in ratio to the value over the whole trial for a particular pair and condition.

Preprocessing DV2

- Minimal-redundancy-maximal-relevance criterion (mRMR)
  - Maximizes feature’s relevance in terms of mutual information between target variable and features while discarding redundant features (Peng, 2007)

➡ **Saccade rate**

Results - Language

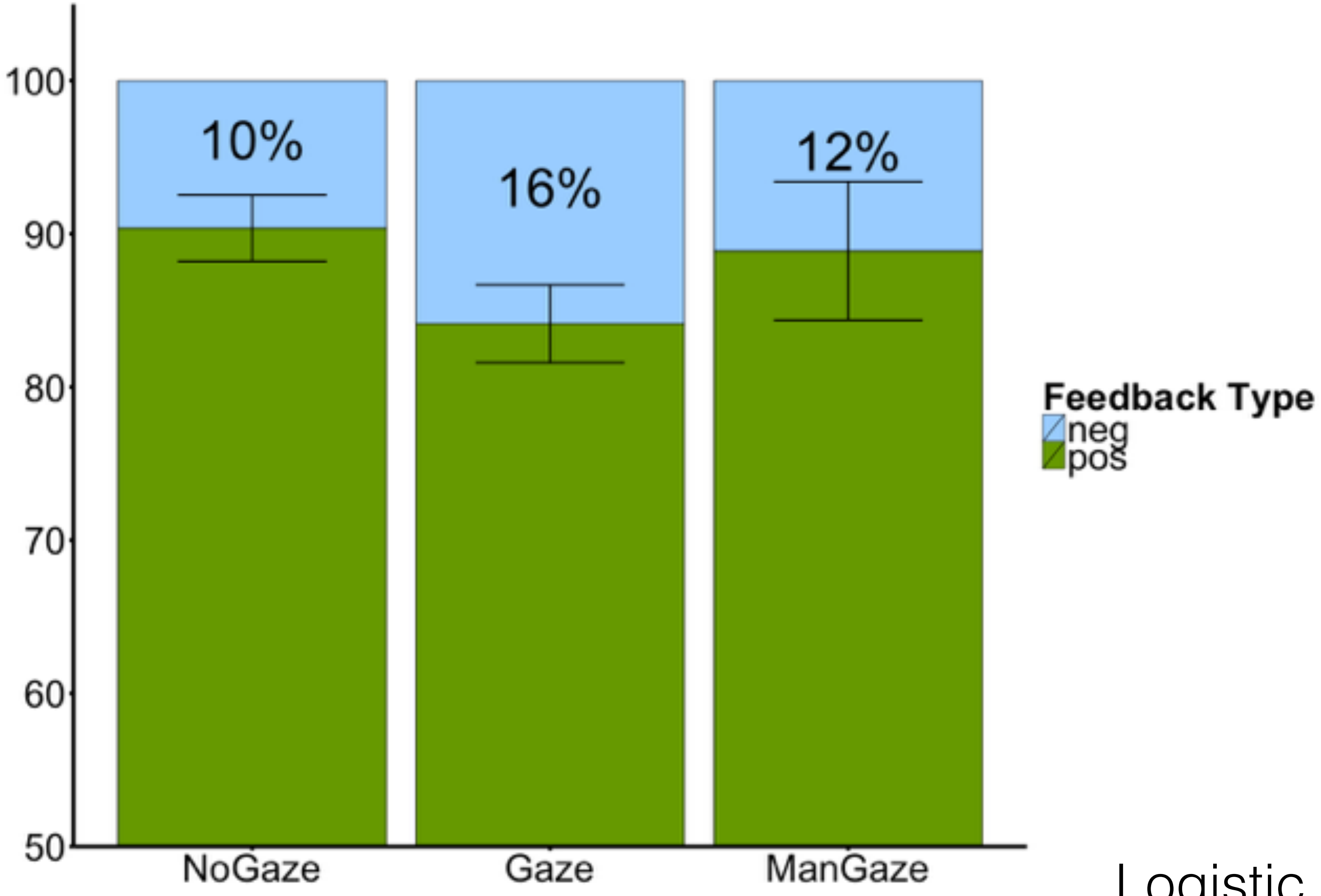
- Performance
  - Success rate ✗
  - Trial duration ✗

Ceiling

- Language
  - No. of spoken words ✗
  - No. of feedback instances ✓
  - Feedback style ?

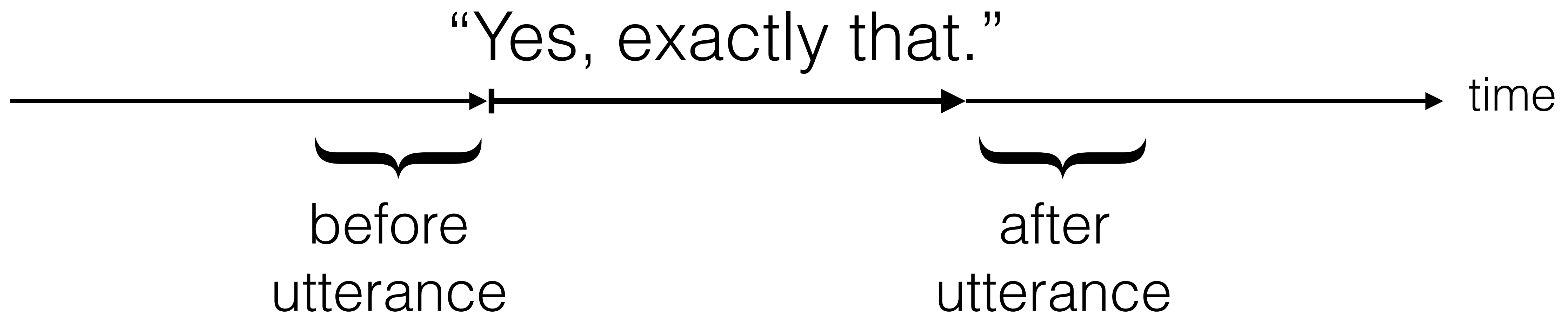
Instruction change

Results - Feedback



Logistic Regression:
Marg.signifiant

Low-level eye-movements



Feedback ***Presence***

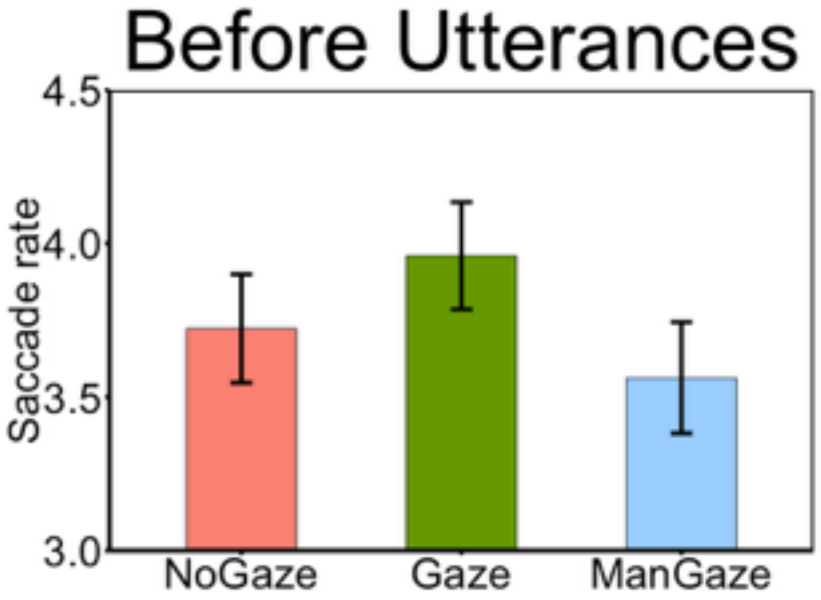
Results - Eye-movements

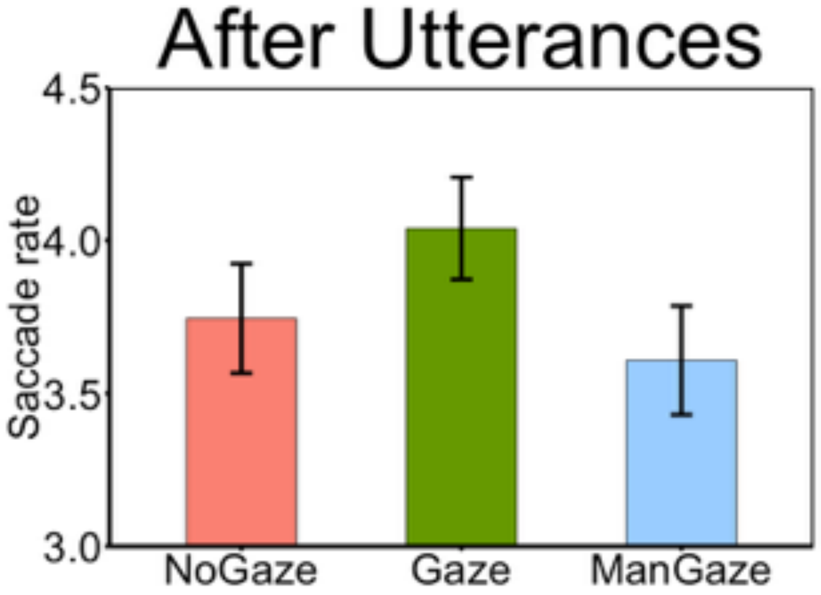
- Effect of *UtterancePresence* on saccade rate *(task recognition)*
  - No effect of *GazeAvailability*

symptom

- Effect of *FeedbackPresence*
  - Interaction with *GazeAvailability*
- Effect of *GazeAvailability* on saccade rate before & after utterances

signal?





Eye-movements & Feedback

- **Manual** annotation of fixations (to target/distractors) up to **5 sec prior to feedback onset**
- No effect of GazeAvailability on patterns found

➡ Feedback timing independent of listener gaze?

Interim summary III

- Instructions change slightly when listener gaze is available
  - More negative feedback
  - But no measurable effect on performance
- Feedback difficult to categorize
- Eye-movement patterns reflect speech process AND change with GazeAvailability
- Gaze-Feedback pattern constant across conditions

symptom & signal

Conclusion

- Listeners follow speaker gaze (and arrows) and form predictions about upcoming referents
    - Difference in strategic use of these cues
- Listeners follow speech & these gaze cues can be exploited by the speaker
    - Efficient use by NLG system
    - Little benefit for real speaker
        - Ceiling, Unnatural situation

References

- Hanna, J. & Brennan, S. (2007). JML
- Staudte, M. & Crocker, M. (2011). Cognition.
- Staudte, M., Crocker, M., Heloir, A., & Kipp, M. (2014). Cognition
- Garoufi, K., Staudte, M., Koller, A., & Crocker, M. (2015). Cognitive Science, in press.
- Koleva, N., Hoppe, S., Staudte, M., & Bulling, A. (2015). Proc. of the Annual Meeting of the Cognitive Science Society, Los Angeles.