Multimodal Interaction

On Attention and Intention

6 May 2010

How do we communicate?

- 1. Speech
- 2. Non-Verbal Sounds
- 3. Body Posture
- 4. Facial Expressions
- 5. Gaze Direction
- 6. Gesture

Speech

- Explicit: “Mary and John have an appointment at 2pm on May 2nd.”
- Vague: “This is too small.”
- General: “Mary likes cake.”
- Pragmatic: “I am cold” -> *“Please close the window.”*

Situated Speech

- Situated: (embodied) Speaker, Listener, environment, context
- Implicit speaker/listener's non-verbal signals
  - Unconscious?
- Always present
- Extremely rich (emotions, attitude, attention...)
- Situating and augmenting speech

Situated Speech

- Spoken language and environment provide huge amounts of information simultaneously

- Processing needs to be fast!

- Using one to facilitate processing the other:

  - Visual information (non-verbal cues)

  - Visual (scene) information

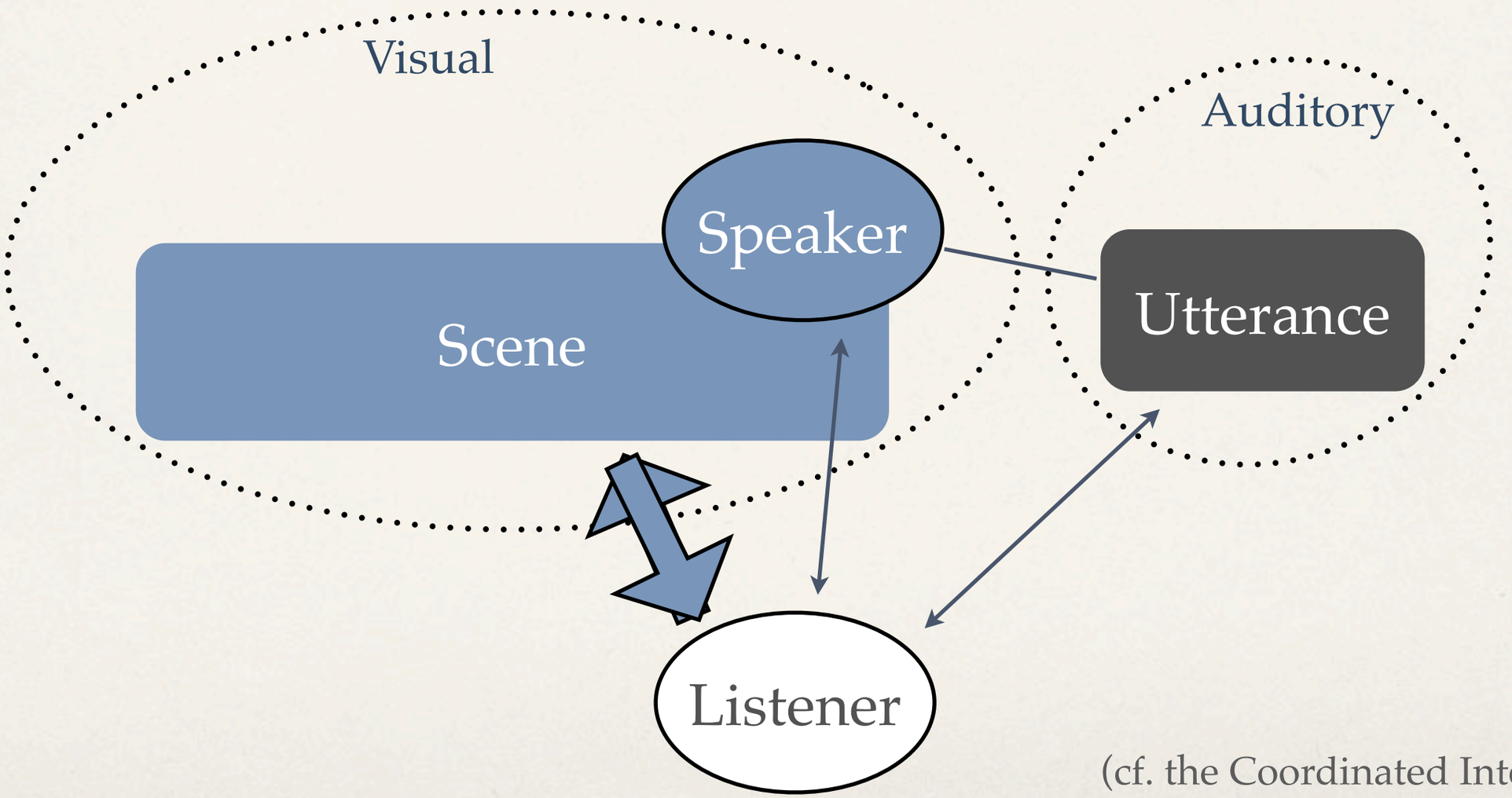
  - Linguistic information

Attention

"Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence."

William James, 1890, "Principles of Psychology"

Situated Speech



(cf. the Coordinated Interplay Account by Knoeferle & Crocker, 2006)

(Visual) Attention

- Attention is a mechanism that changes the mental state of the attending individual

- Directing someone's attention is an important aspect of communication and it :

  - 1. Entails the presence and acknowledgement of mental states

  - 2. Comprises head movement, gaze, gesture, speech as tools

1. Mental States

- Prerequisites:
  - Assigning and understanding mental states (to the partner)
  - Seeing goals and intentions in the partner's actions
  - Understanding that one can influence others' mental states

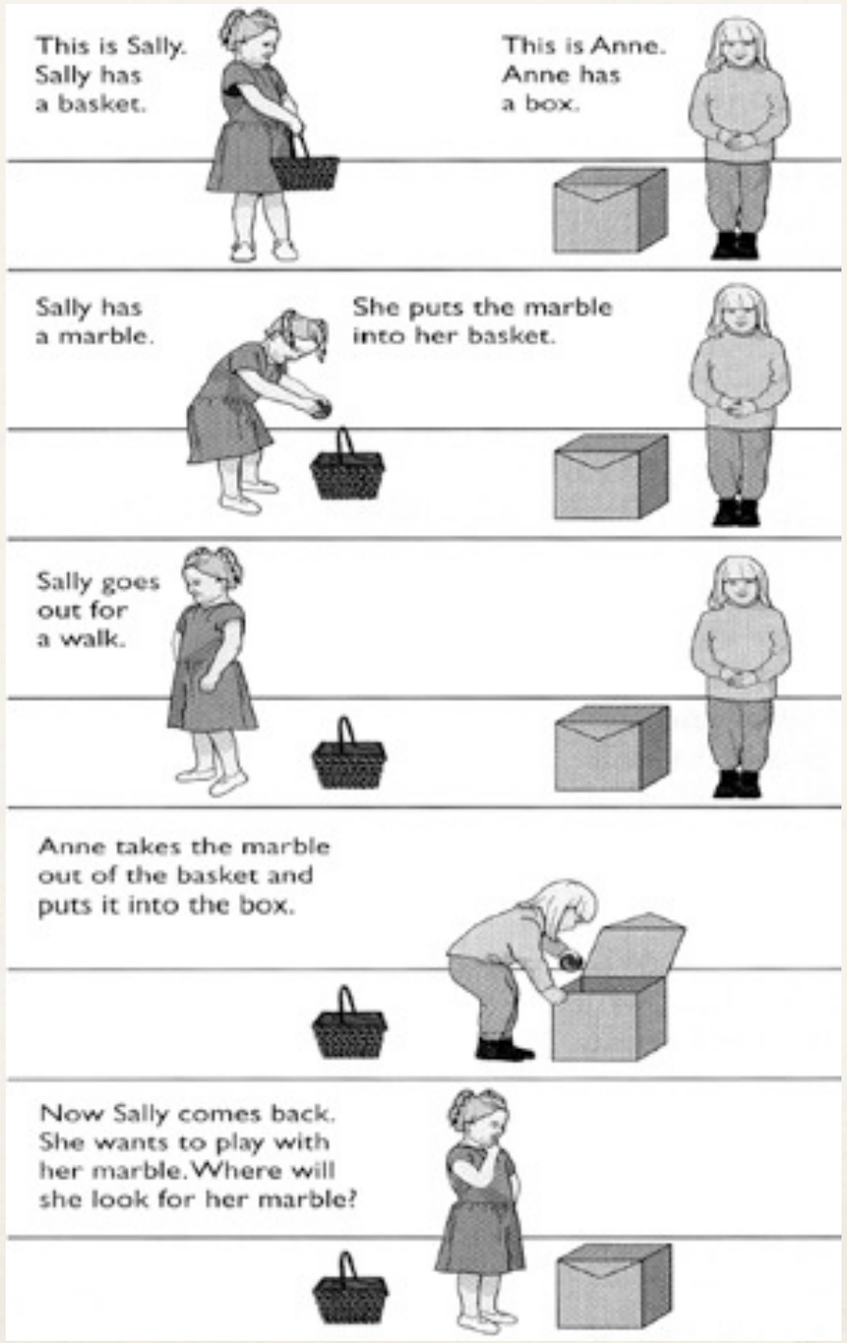
→ Theory of Mind (Premack & Woodruff 1978, Baron-Cohen 1995)

  - To have a theory of mind means to use knowledge about mental states, and about epistemic mental states (believing, knowing, pretending) in particular, in a theory-like way.

1. Mental States (ToM)

- I.e., a theory of mind (ToM) allows us to:
  - Take someone else's perspective,
  - Understand that *seeing* means *attending and perceiving* means *knowing*,
  - Infer and manipulate partner's epistemic states
    - Make her aware of something; deceive her

1. Mental States (ToM)



(Pratt & Bryant 1990, Baron-Cohen 1995)

1. Mental States (ToM)

Conditions:

- **Desire**: Which one do you want?
- **Goal**: Which one will you take?
- **Reference**: Which one is the *beb*?

Charly:

- Which one does C. want?
- Which one will C. take?
- C. says “There’s the beb!”. Which one does C. say is the *beb*?

(Baron-Cohen et al. 1995)

Visual Attention

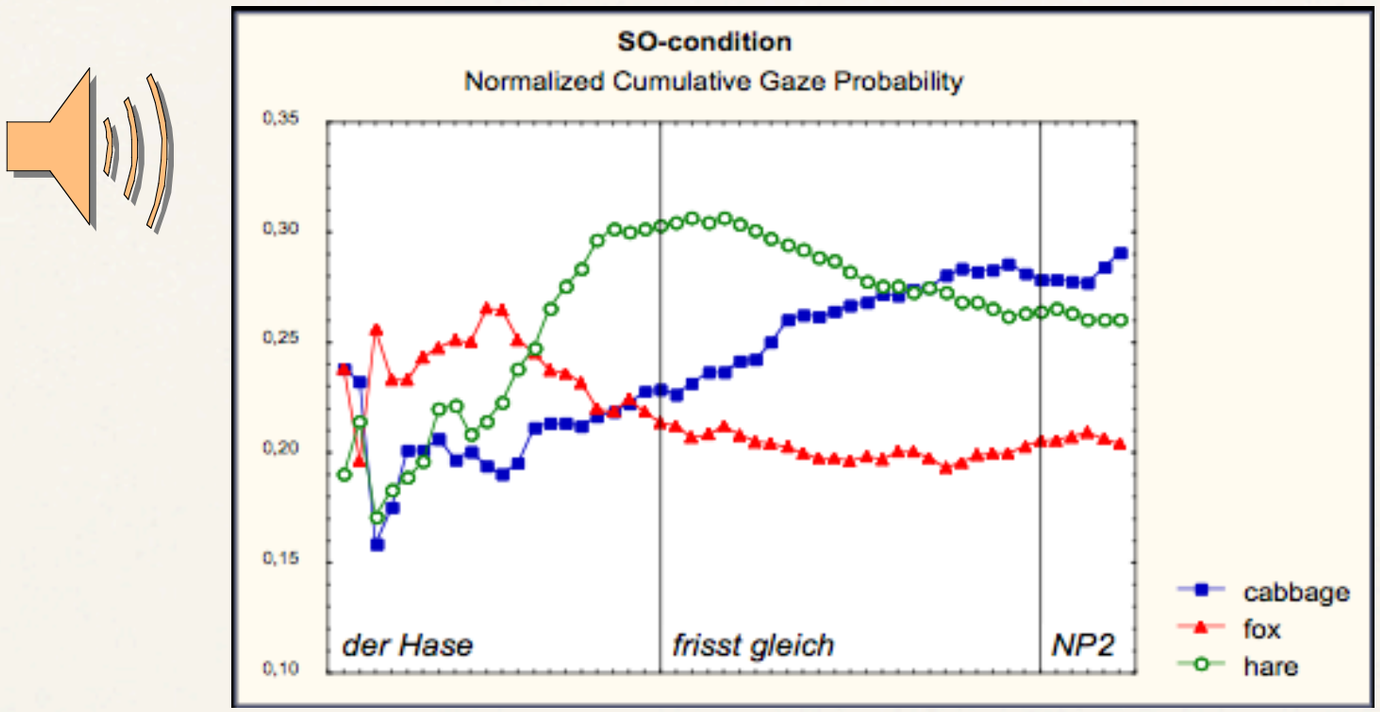
Remember:

- Directing someone's attention is an important aspect of communication and it :
  - 1. Entails the presence and acknowledgement of mental states
  - 2. Comprises head movement, gaze, gesture, speech as tools

2. Attentional Cues

- **Gaze / Seeing:**
  - 1. is part of controlling one's own visual attention, and
  - 2. is - by expressing the attentional focus - also a direction giving cue.
- Visual World Paradigm: Studies relying on and investigating eye movements during language comprehension/production as a cue to what is being processed and when
- Interaction Studies: Study gaze as additional cue directing others' attention

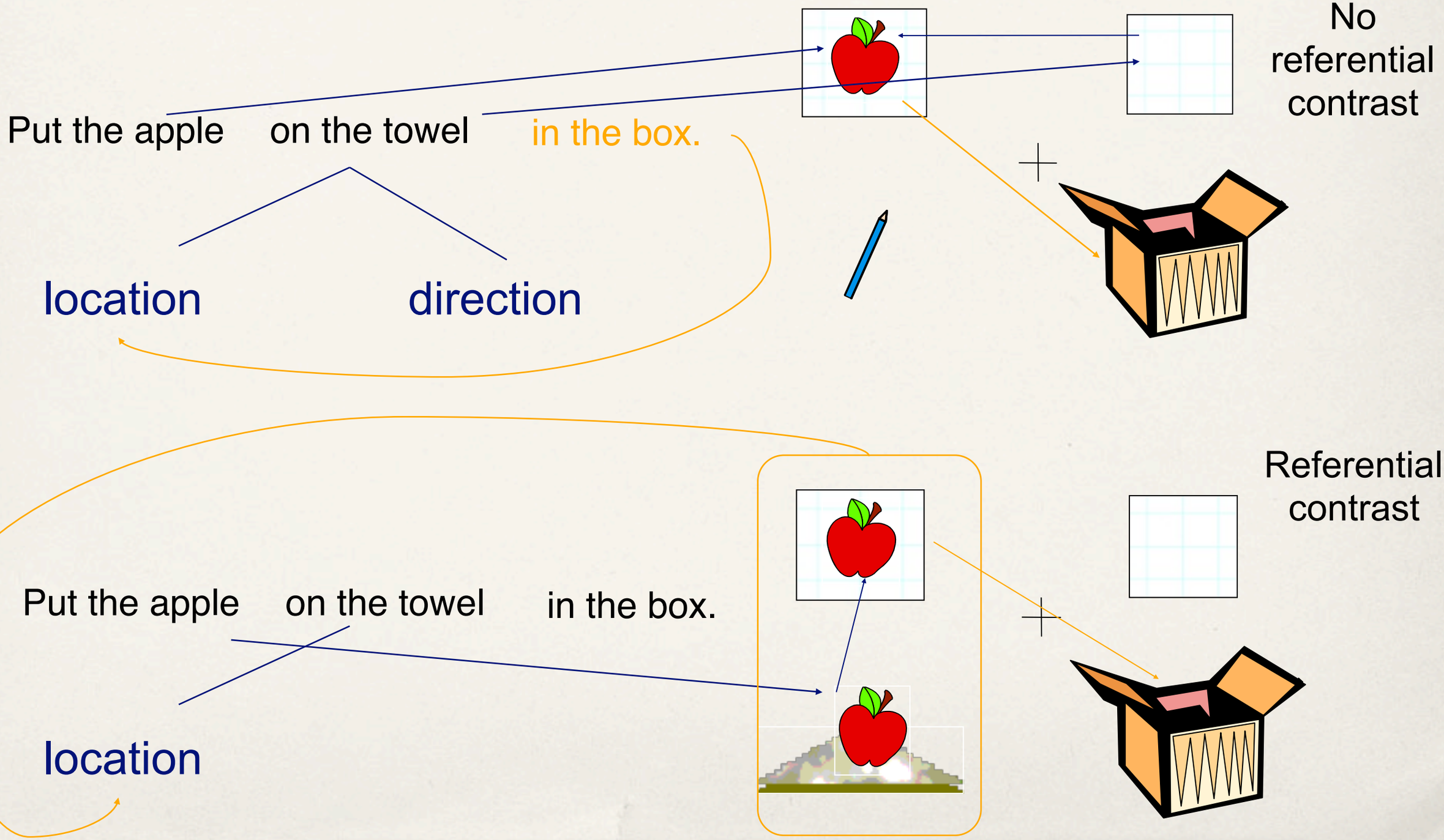
Eye-tracking in scenes



- Attention to objects in the scene is closely time-locked to comprehension
  - Makes it possible to use eye-tracking in scenes during utterance presentation to investigate spoken comprehension
  - Permits us to examine use of scene information for comprehension

Visual World Studies

Tanenhaus et al. 1995

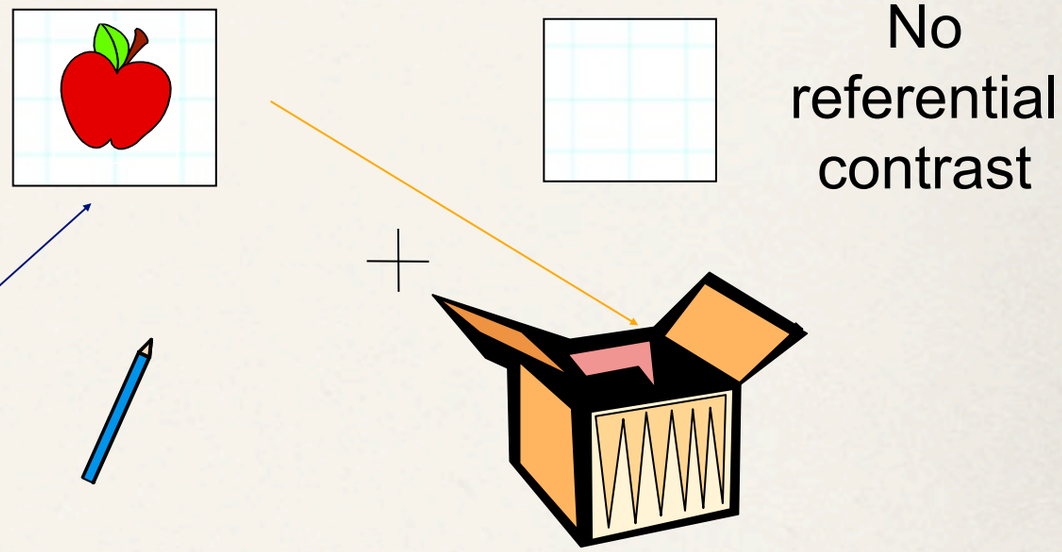


Visual World Studies

Tanenhaus et al. 1995

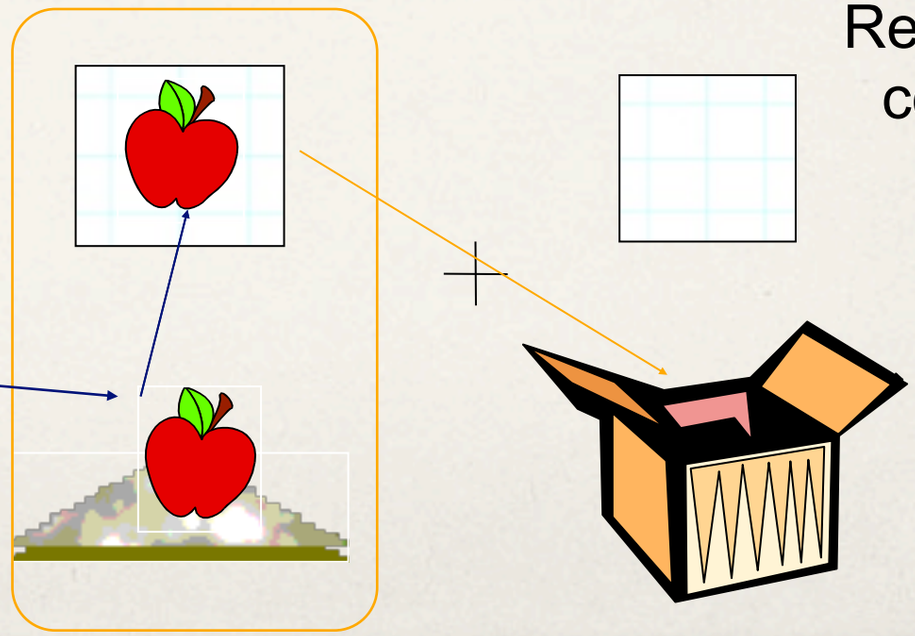
- What are the effects of the 2-apple scene?
- Establishes contrast between 2 objects: apples
- This referential contrast enables structural disambiguation
- Why do we know this?
- Because there are no looks to the target-towel for the referential-contrast condition (there are such looks in the "no-referential contrast" condition)
- And because there was another "control-condition" where the sentence was unambiguous : "Put the apple **that's** on the towel in the box.

For that "control-condition" the pattern of eye-movements to objects was in both types of contexts (1-apple, 2 apples) the same as for the ambiguous sentences in the 2-apples context:



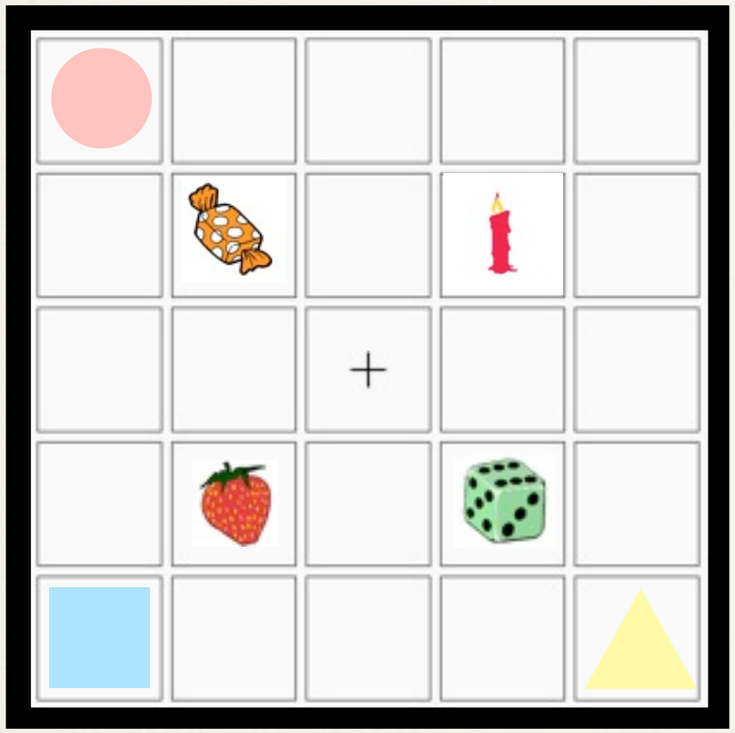
No referential contrast

Put the apple that's on the towel in the box.

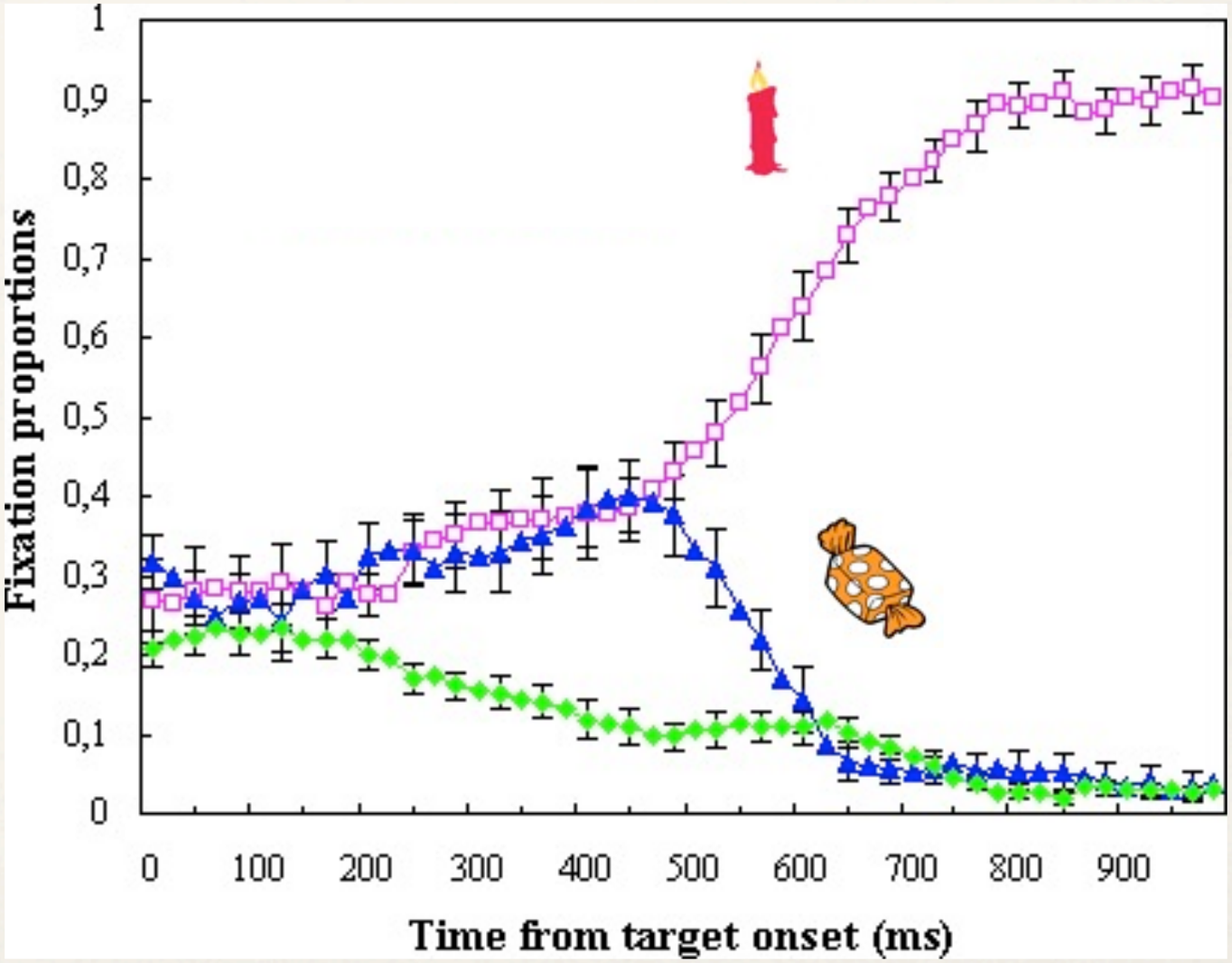


Referential contrast

Lexical access over time



"Pick up the *candle*"



Incremental Semantic Interpretation

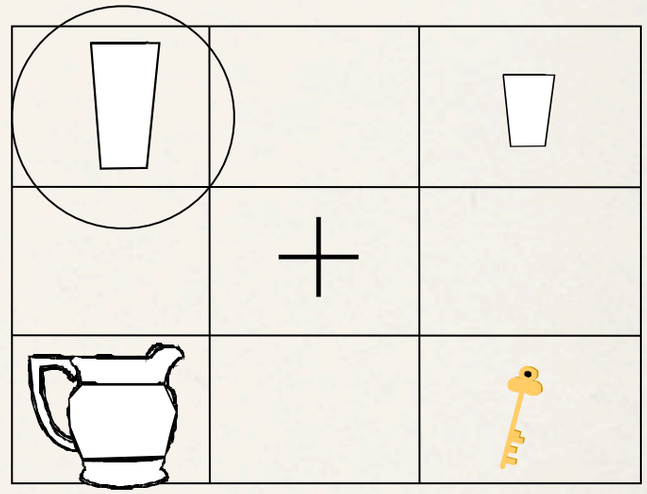
Sedivy et al. 1999

More visual referential ambiguity

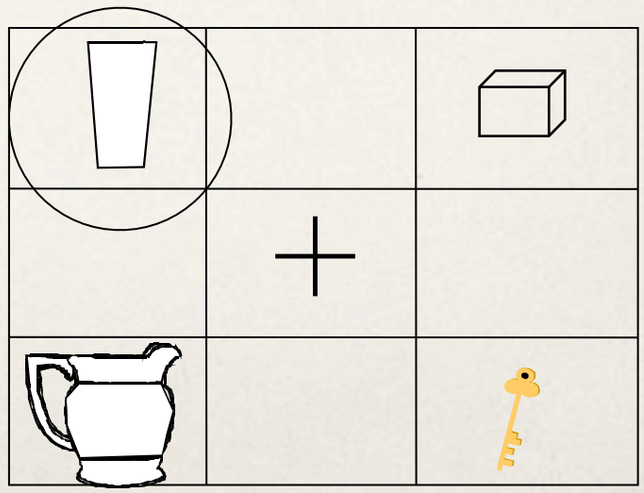
- Influence of **visual contexts** on
  - **determination of reference to entities**
  - **Properties of objects (small, tall)**

Pick up the tall glass and put it below the pitcher.

Two same-type objects
that differ in 1 property: size



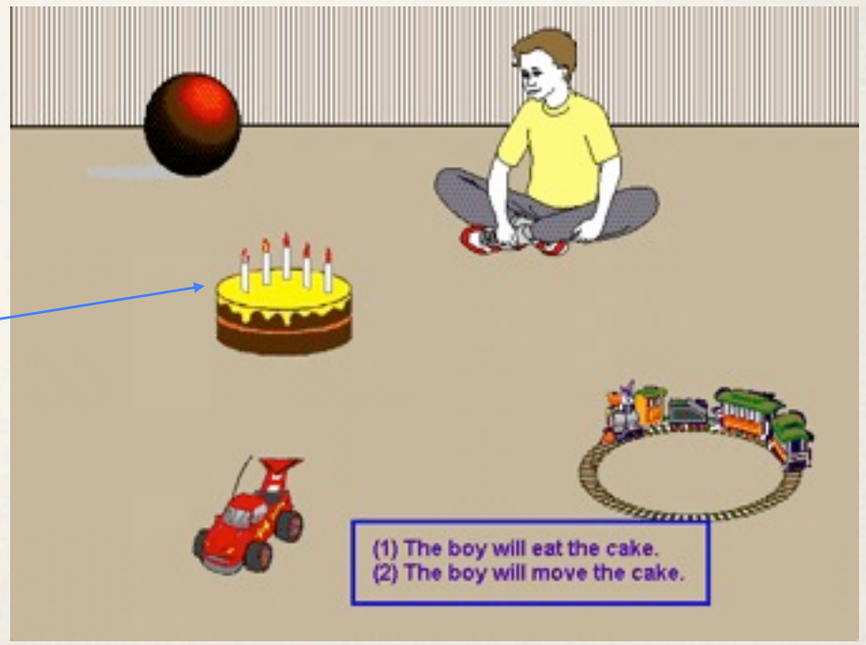
No contrastive objects
of the same type



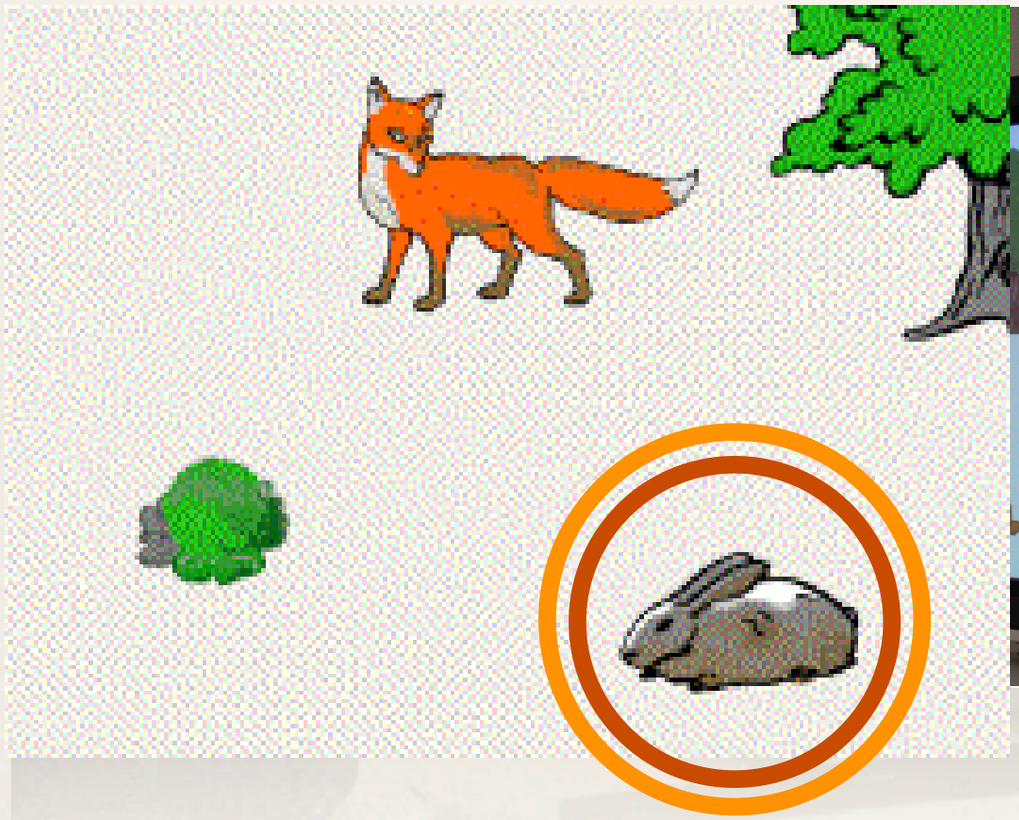
- More rapid looks to the tall glass before hearing "glass" in the contrastive than non-contrastive condition

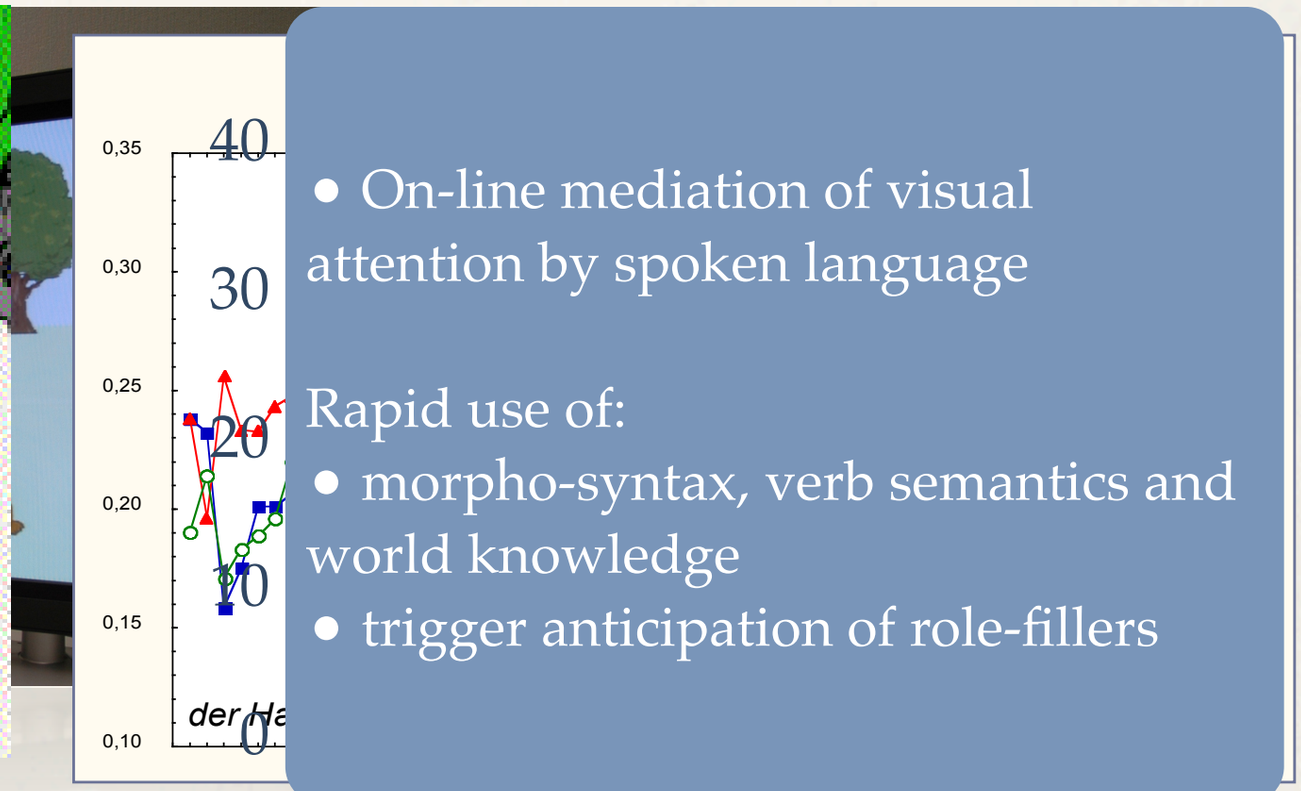
Anticipatory eye-movements

- "Eye-movements to an object in a scene before it has been named"
- Do verb selectional restrictions allow anticipation of as yet unmentioned postverbal argument/ its referent in the scene?
  - Verb selectional restrictions: *eat* can take only edible objects as arguments
- What is anticipated?
- Why is an object anticipated?
- "The boy will move the cake."
  - train, ball, toy car and cake can be moved
- "The boy will eat the cake."
  - highly restrictive: only the cake is edible



Anticipation in Visual Worlds





- On-line mediation of visual attention by spoken language

Rapid use of:

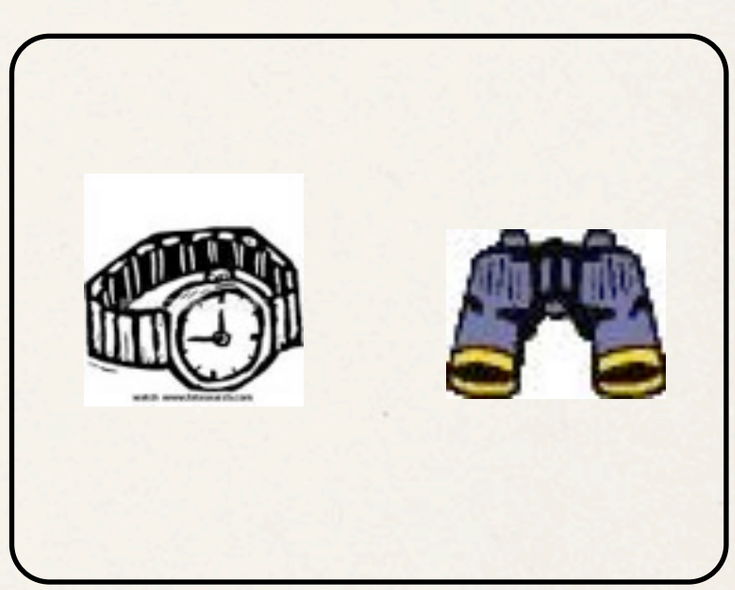
- morpho-syntax, verb semantics and world knowledge
- trigger anticipation of role-fillers

				Patient	Agent
SVO	Der Hase *The hare (nom)*	frisst *eats*	gleich *soon*	den Kohl *the cabbage (acc)*	
OVS	Den Hasen *The hare (acc)*	frisst *eats*	gleich *soon*	der Fuchs *the fox (nom)*	

Kamide et al, 2003.

Eye-movements during production

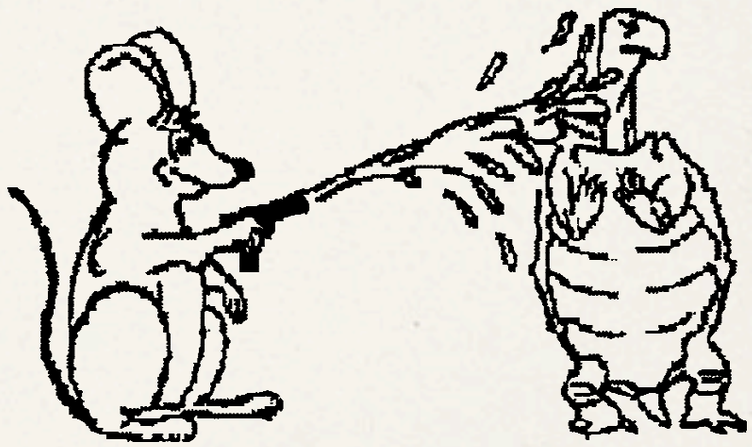
- Manipulations: Frequency, Ease of Identification (contours)
- Naming Task
- Categorization Task
- Naming Latencies and Viewing Time affected only during lexical access
- What does this say about the role of viewing time?



Meyer et al. 1999

Eye-movements during production

- Description/Naming Task
- Subject is inspected most before speech onset, patient is inspected most during speech
- More precisely: Eye-movements to individual objects 800ms-1000ms prior to mentioning



Griffin & Bock 2000, Griffin 2001

Eye-movements during production

- Eye-movements to investigate time course of word selection during sentence production
- Descriptions of object arrangements
- Manipulations: Frequency, Codability
- "The A and the B are above the C."
- Viewing time affected by frequency and codability
- But: Only for individual object! Word selection is obviously done "on the fly".

Griffin & Bock 2000, Griffin 2001

Summary Visual World Studies

- Which kinds of information may influence spoken sentence **comprehension** ?
- Incremental use of
  - Linguistic knowledge
    - Verb selectional restrictions
    - Scalar adjectives
    - Case-marking + verb plausibility
  - Visual scene information
    - Properties of objects (size, shape, texture)
    - Referential contrast between objects
    - … ? … well, how about events?

- Referential visual contrast: structural disambiguation
- Adjectives: incremental semantic interpretation
- Case-marking & verb plausibility: thematic role-assignment

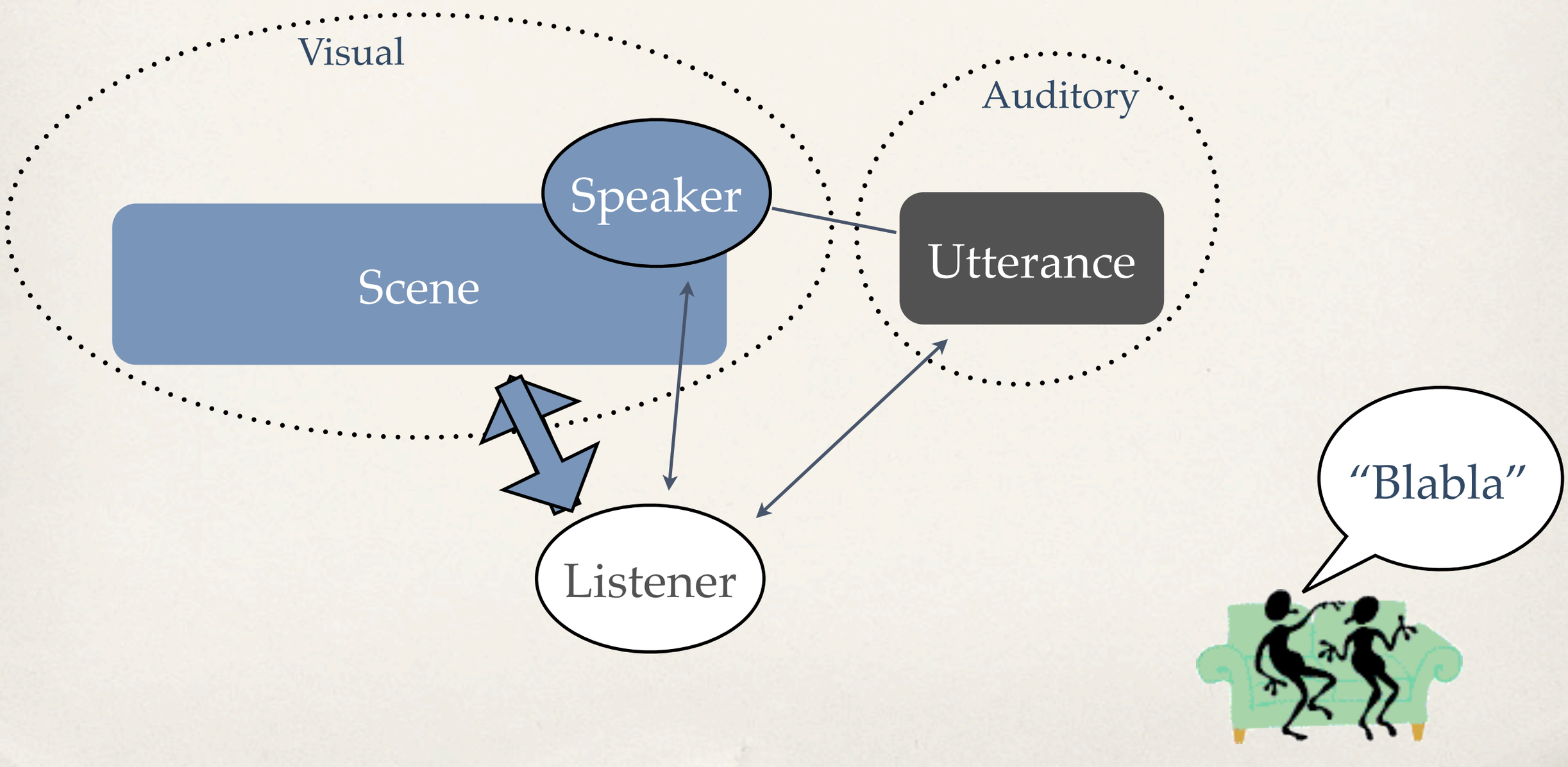
} identify scene objects

Summary Visual World Studies

- Which kinds of information may influence sentence **production** ?
- Incremental use of visual scene information
    - Properties of objects (ease of identification)
- Lexical accessibility (frequency, codability)
- What is the link between viewing times (visual information) and naming??

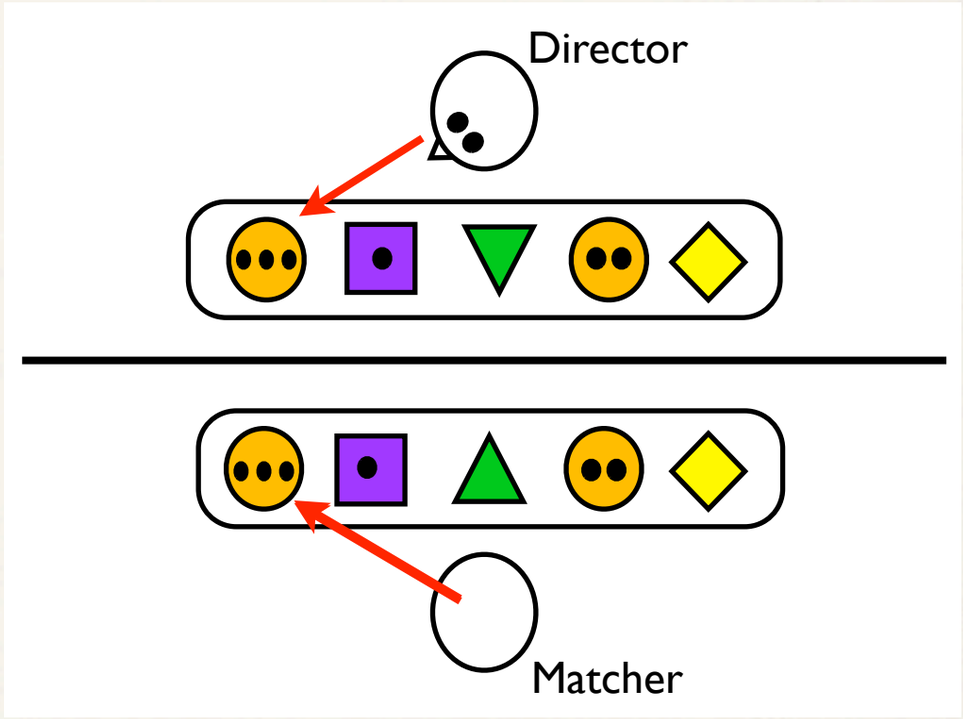
Situated Speech

Remember:



Eye-Movements in Situated Speech

- Listener can see speaker's gaze
  - "Move the circle with ...
  - ".. three dots to location A."
- Speaker can see listener's gaze
  - "Move the circle with ...
  - ".. yeah, that, to location A."
- How useful is this?
- How does this affect language comprehension & production?

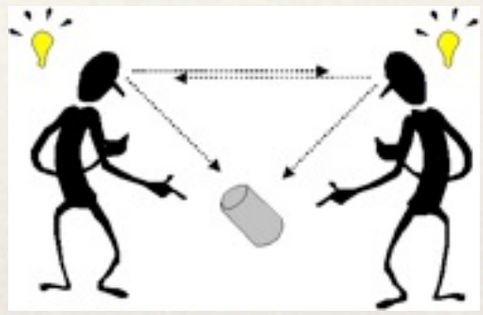


Hanna & Brennan 2007, Clark & Krych 2004

Attention & Language

Summary:

- People look at what they hear (comprehension)
- People look at what they say (production)
- People look at where other people look



Eye-Movements & Language

- Attention and Joint Attention
  - What are the prerequisites? What processes are involved? What are the results?
- Which tools/cues can be used to direct attention?
  - When and why are different cues used? How do they interact? Are they used intentionally or unconsciously?
- How does the look and use of such a tool/cue affect its influence and perception? Could also affect whether joint attention is/can be established?

Practical Matters…

- How to read a paper!?

Title, Introduction, Conclusions

What is the paper about?

1. Description of a phenomenon or problem,

2. and formulating a research question.

3. Summarising state-of-the-art research in the area, addressing this question.

4. Ending with the hypothesis the paper wants confirm.

Introduction: Nass & Moon

1. People know that computers are not humans, and yet in interaction with computers they behave as if that's not the case.

2. Why is this so? What causes people to assign, e.g., social categories to computers?

3. Explaining and ruling out alternative explanations: anthropomorphism; intentional orientation to the programmer; demand characteristics

4. People "mindlessly apply social rules and expectations to computers".

Methods

- Hypothesis, Argumentation chain
- Empirical Evidence
  - Design? Conclusions valid?
- Are other explanations for the presented results possible?

Methods: Nass & Moon

- Hypothesis: Social scripts are applied in HCI that are "inappropriate" -> mindless
- Presuppositions: **Mindless scripts elicited when computer shows "enough" but not entirely human cues**
- Empirical evidence:

1. Gender stereotypes are applied
2. Proposed ethnicity triggers typical reactions
3. Team membership
4. Politeness, no feelings hurt
5. Reciprocity, "returning the favour"
6. Premature cognitive commitment to expertise
7. Personality traits

one should ask oneself:
what ARE the cues that elicit these responses?

human-like computer cues:
-words,
- interactivity
- filling of traditional human roles

- Design:
  - **Instructions?** (check 3., asked to focus on individual responsibility)
  - Does the measure capture the proposed effect? (check 4., adhering to expected behaviour?)
  - Do manipulations facilitate this (check 7. computer's use of language)

Discussion, Conclusion

- Summarising results.
- Addressing flaws,
- And alternative explanations.
- Completing the argumentation chain.
- Questions raised, outlook.

Discussion, Conclusion: Nass & Moon

- Linking alternatives to “thoughtful” application of social rules
- Rejecting alternative explanations:
  - Anthropomorphism: Adults explicitly denying social treatment of computers/machines
  - Intentional orientation to the programmer: Explicit denial, same programmer, null effect for manipulation of term usage “computer” vs “programmer”
  - Demand characteristics: It was never pretended that the computer was an individual, and subjects were unaware of their behaviour.

Discussion, Conclusion: Nass & Moon

- Questions Raised:
  - When and why does mindlessness occur?
  - More human features -> *even more* social responses?
  - Would errors, that are not typical for humans, remind people of the "nonhumanness"?
  - How about combining a very human-like module with a crude, non-human module...
  - Comparison with human-human interaction studies useful? What would we learn?

Discussion, Conclusion: Nass & Moon

- Conclusions and Outlook:

1. Behaviours that are controlled by more primitive and automatic processes are more likely to be mindlessly elicited than more socially constructed behaviours.

2. Rules that are used frequently are more likely to be mindlessly elicited than rules that are used rarely.

3. Social behaviours that are uniquely directed at members of a person's culture may be more difficult to elicit via computers - since the computer may more often remind the user of its non-membership

Conclusion

- Presented a process that accounts for seemingly bizarre responses to computers
  - Agree or not?

Questions

"This concludes my lecture on non-verbal communication. Any comments or questions?"