Specialization Module

Speech Technology

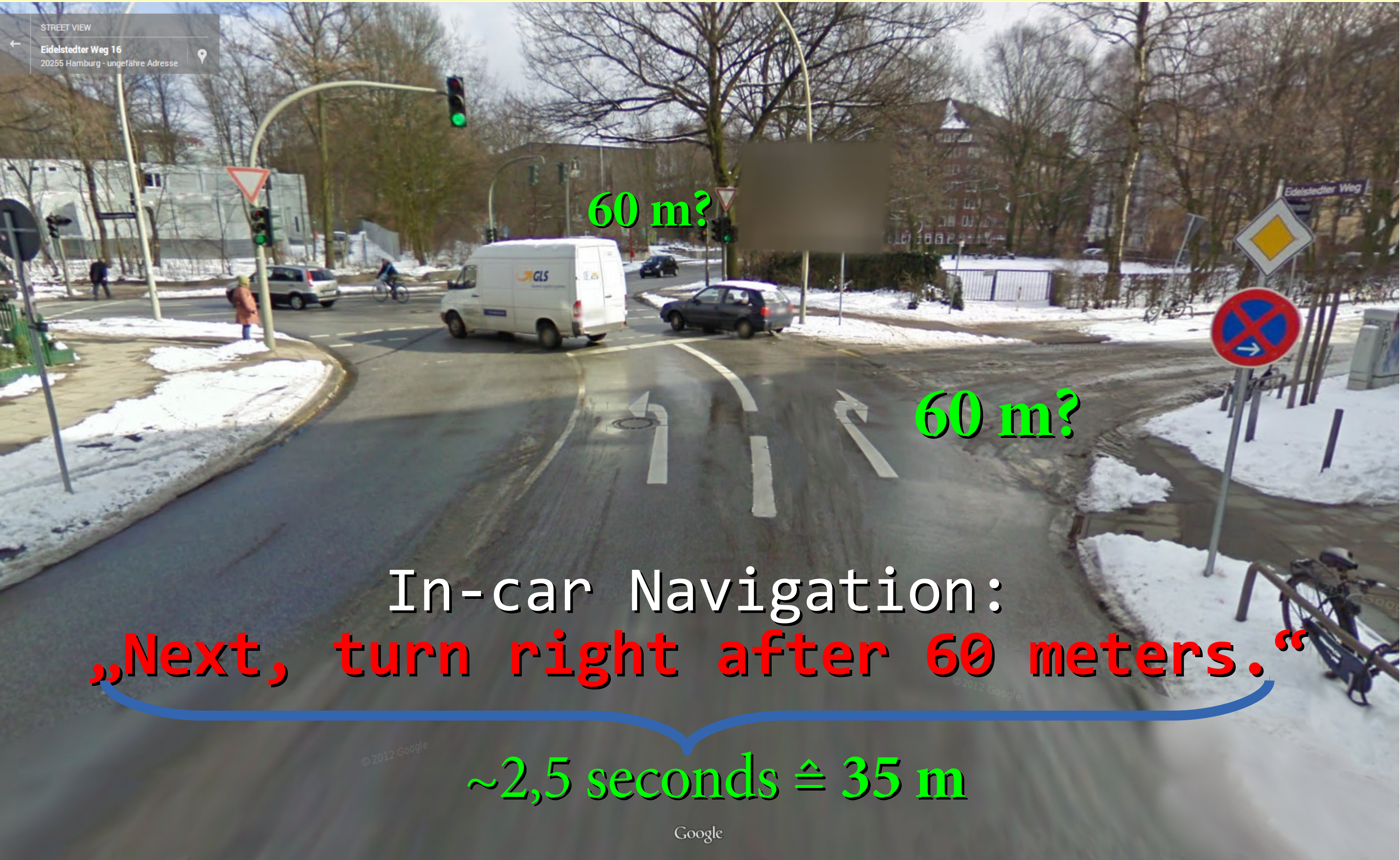
Timo Baumann
baumann@informatik.uni-hamburg.de

Universität Hamburg, Department of Informatics
Natural Language Systems Group

Incremental Processing

What's wrong with conventional interactive spoken language processing systems?

Example



60 m?

60 m?

In-car Navigation:

„Next, turn right after 60 meters.“

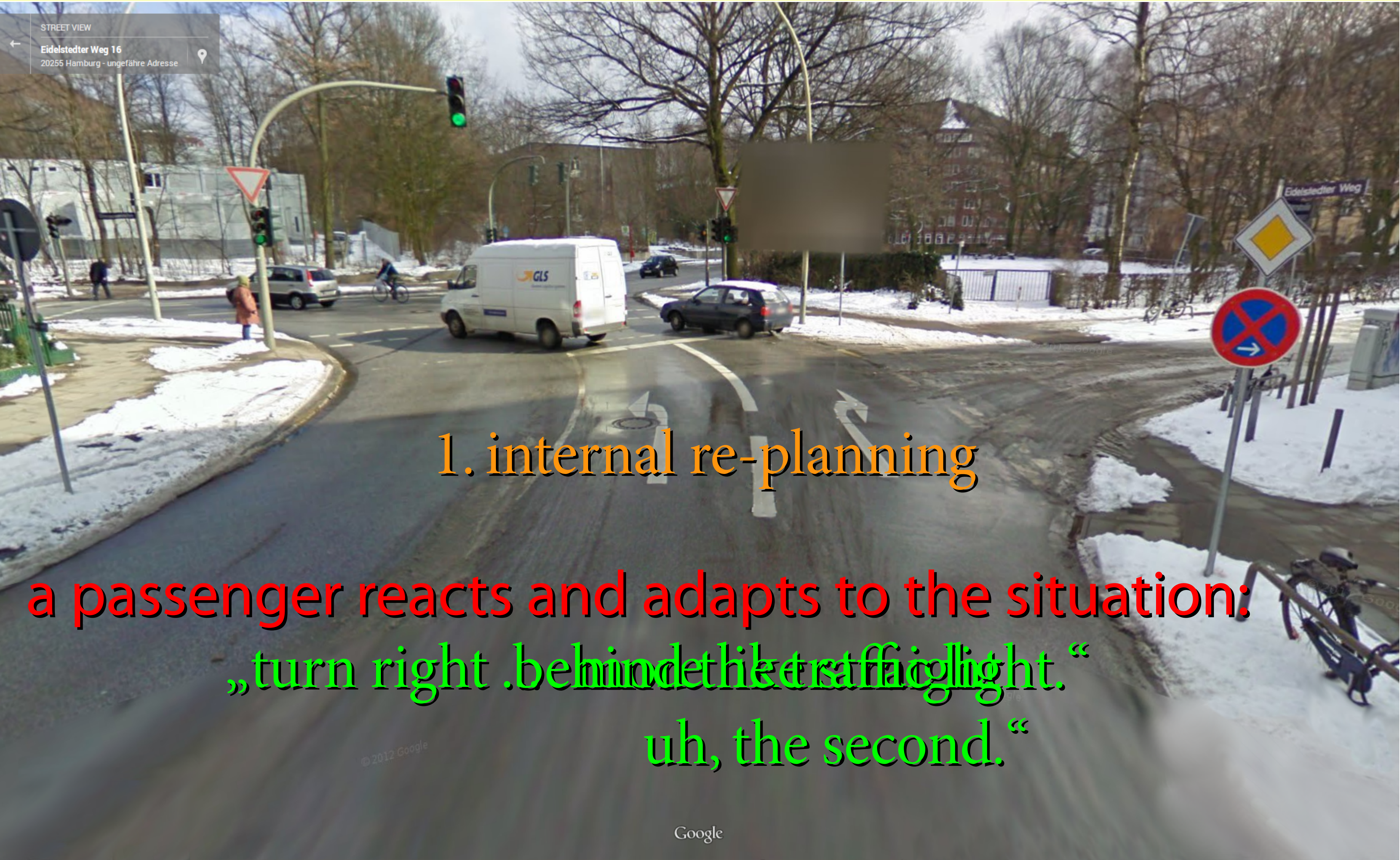
~2,5 seconds ≙ 35 m

Example

Spoken language unfolds in time

↦ this is both a challenge and the solution

Human speakers are *responsive*.



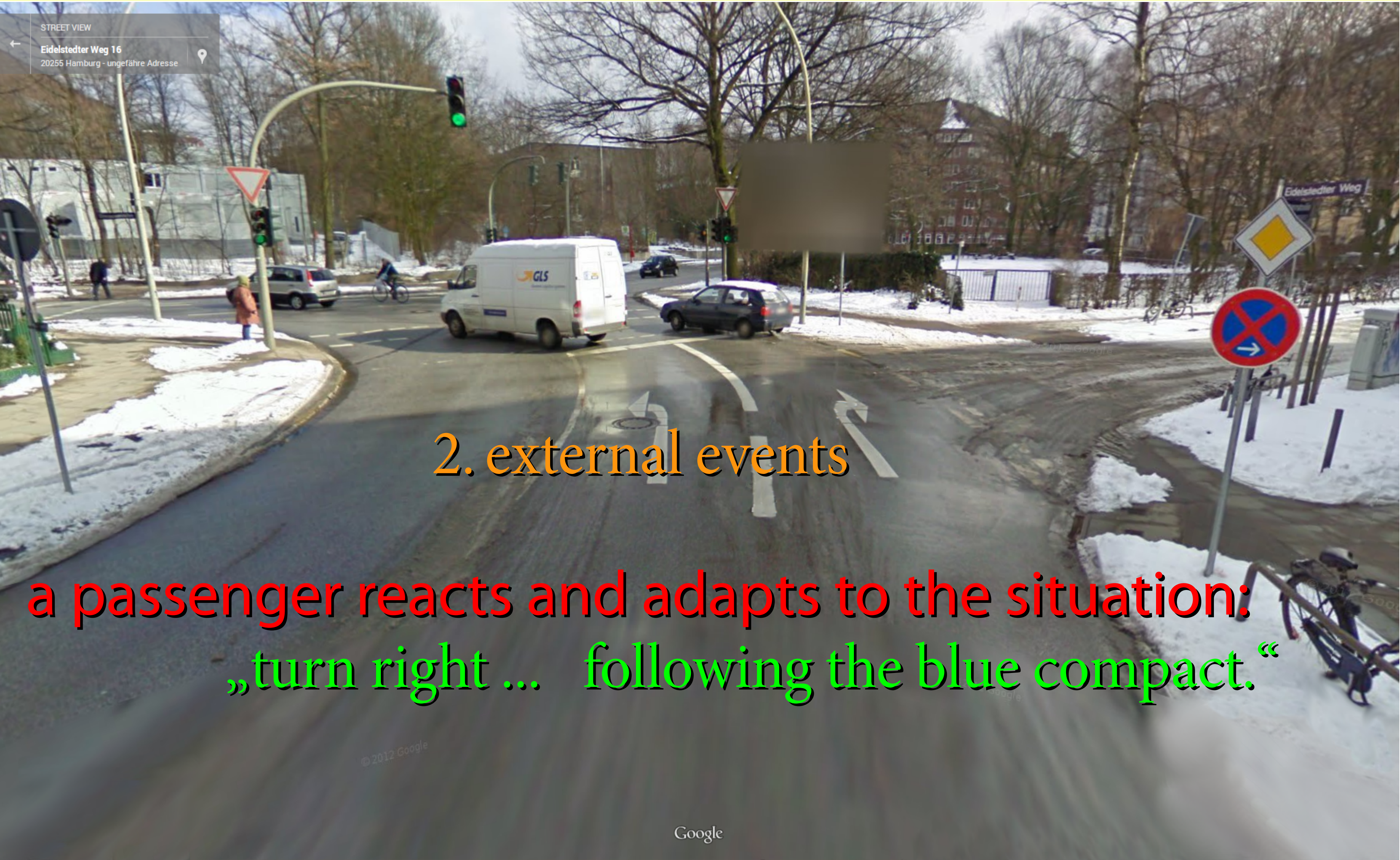
1. internal re-planning

a passenger reacts and adapts to the situation:

„turn right .behind the traffic light.“

uh, the second.“

Human speakers are *responsive*.

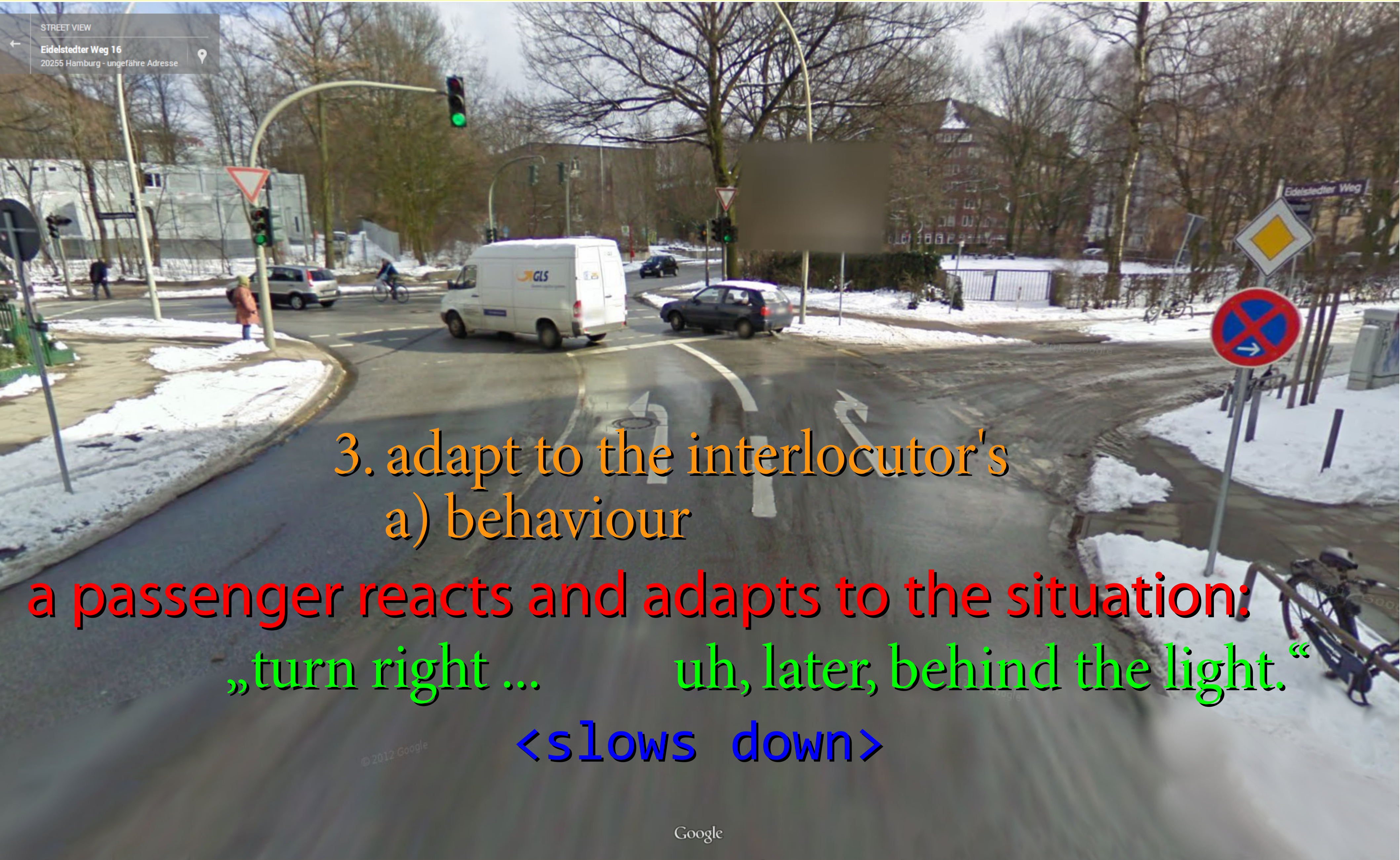


2. external events

a passenger reacts and adapts to the situation:

„turn right ... following the blue compact.“

Human speakers are *responsive*.



3. adapt to the interlocutor's
   a) behaviour

a passenger reacts and adapts to the situation:

„turn right ... uh, later, behind the light.“

<slows down>

Human speakers are *responsive*.

3. adapt to the interlocutor's

b) spoken feedback

a passenger reacts and adapts to the situation:

„turn right ... yes, behind the light.“

„behind the light?“

other scenarios requiring responsive behaviour

- Simultaneous interpreting

  mostly internal re-planning

- Human-robot interaction

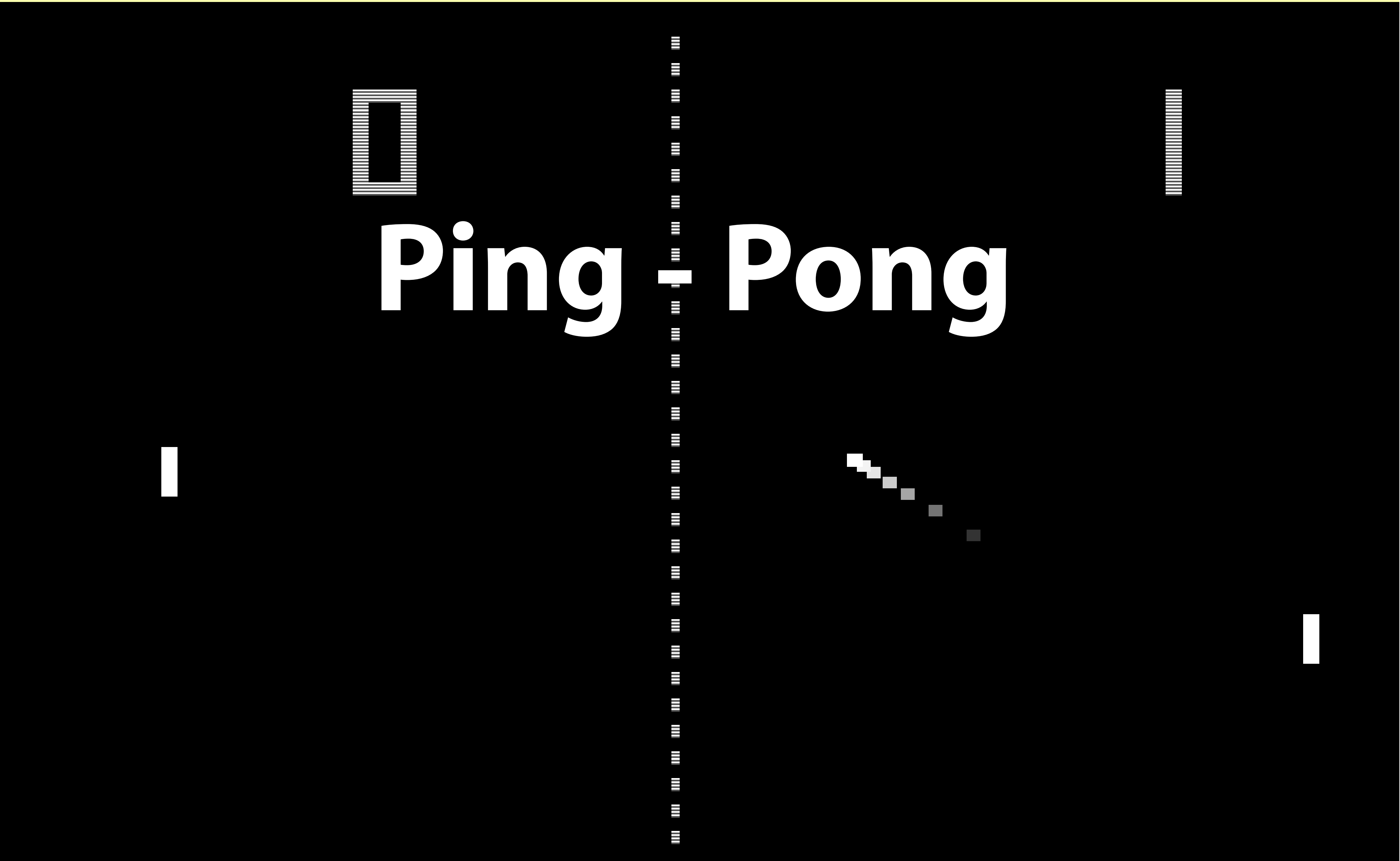
  mostly external events

- Interaction with conversational dialogue systems

  mostly adaptation to user feedback

➔ almost any kind of spoken *interaction* profits from highly responsive behaviour

predominant form of human-computer spoken interaction



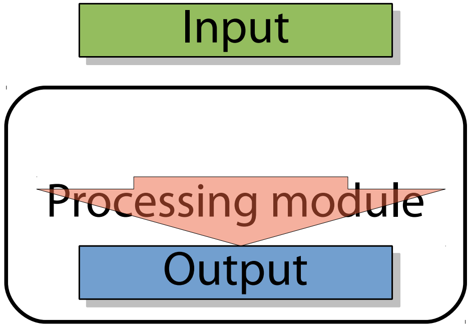
Ping - Pong

Incremental Processing: a Definition

- an incremental processor consumes input and generates output in a piece-meal fashion.
- (preliminary) output is generated before all input has been consumed (at least in some situations).

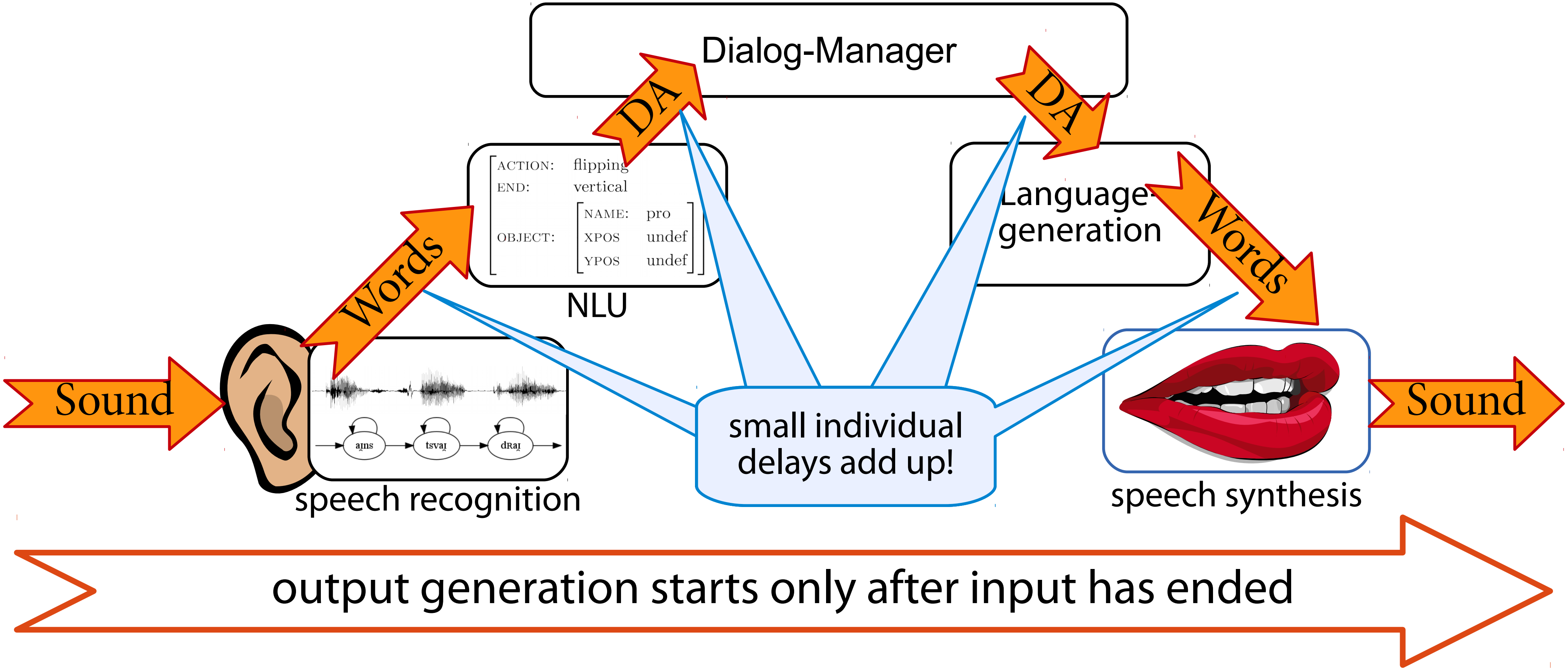
Incremental vs. Non-incremental Processing

- non-incremental, *decoupled* processing:



- Processing is effected after the input → delay!
  - in a modular system: delays add up

a modular dialogue system



Dialog-Manager

NLU

Language-generation

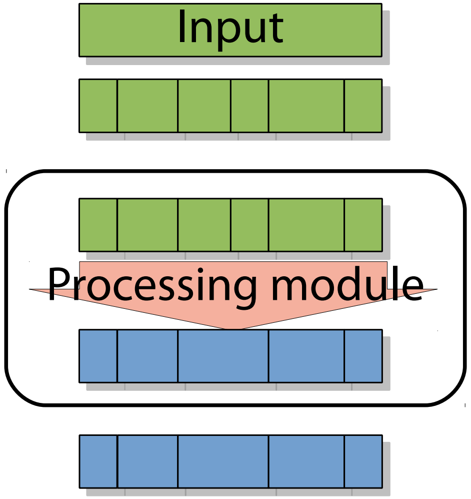
speech recognition

speech synthesis

small individual delays add up!

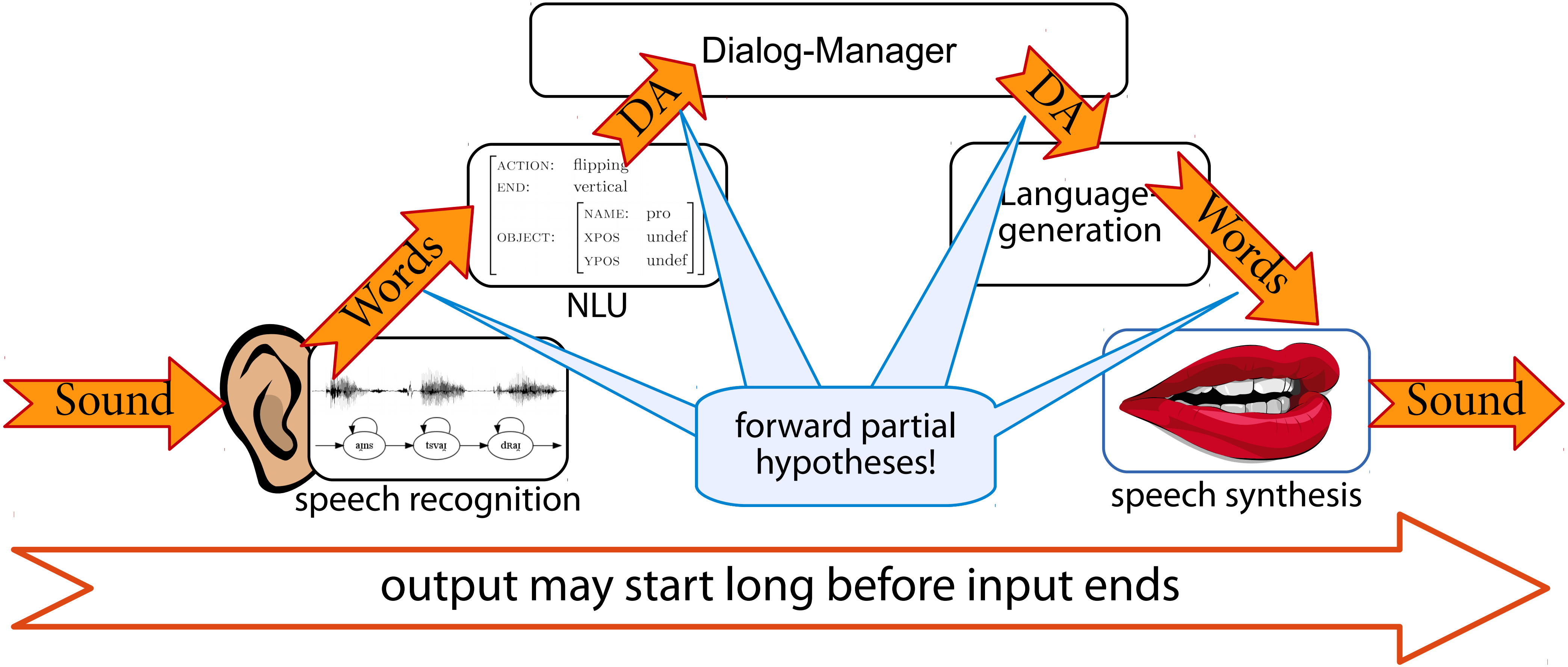
output generation starts only after input has ended

Incremental Processing



- input consists of individual units that are consumed one-by-one (e.g. speech audio, words, ideas, …)
- input is consumed unit-by-unit, and output is generated
- input units may be aggregated to larger units

a modular dialogue system



Sound

Words

DA

Dialog-Manager

DA

Language-generation

Words

Sound

NLU

speech recognition

forward partial hypotheses!

speech synthesis

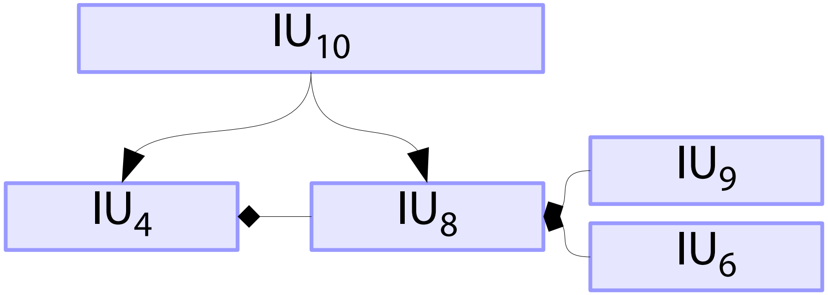
output may start long before input ends

Incremental Processing: Limitations

- hypotheses are based on *what has been seen so far*
  - later input may result in changes
- example speech recognition:
  - input: [f O 6] → this sounds like "four"!
  - addition of [t i:] → together, this sounds like "forty"!
  - what happens if [n] is next? then [EI dZ 6 z]?
- ***limited context* as future input is not considered**
  - either, results will deteriorate, or:
  - allow to **revise previous hypotheses**
    - as a result, the input of following modules is revised,
      which will then also have to reconsider their output and so on

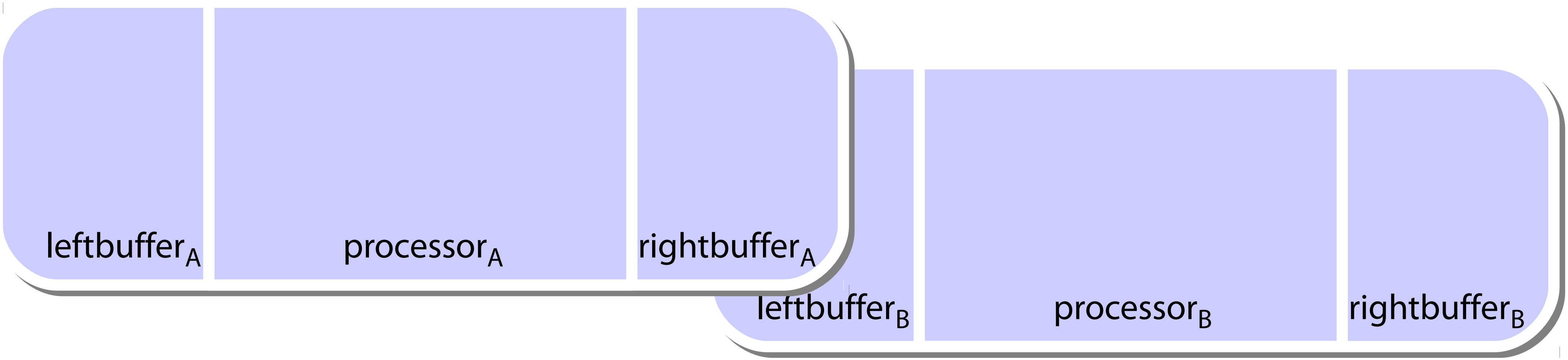
the Incremental Unit

- linked with corresponding unit(s) on the lower/higher levels of abstraction
- linked with neighbouring units on the same level
  - one link pointing backward in time
  - potentially multiple links pointing forward



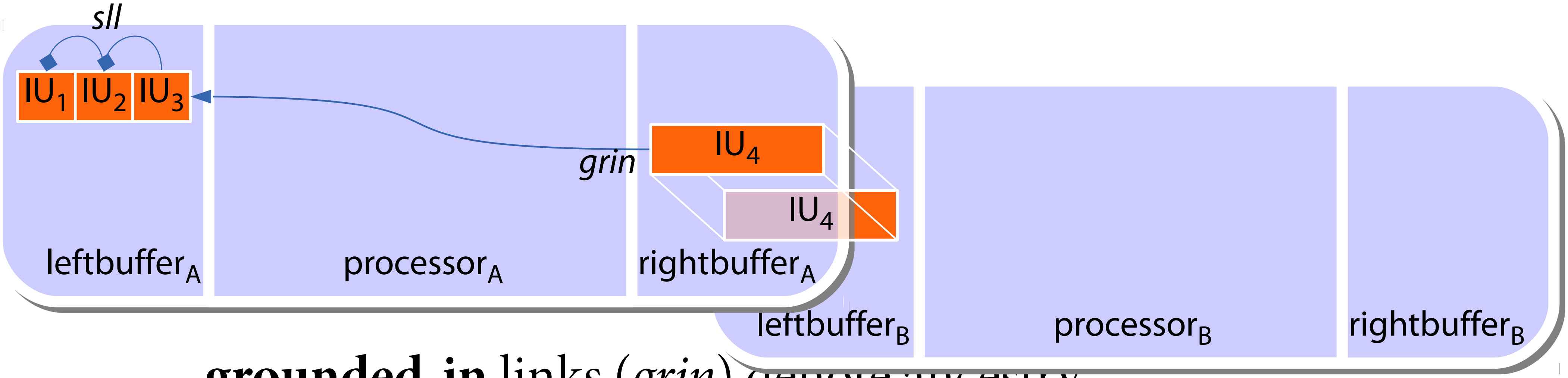
Processing modules

- processing modules are connected via buffers



Processing modules

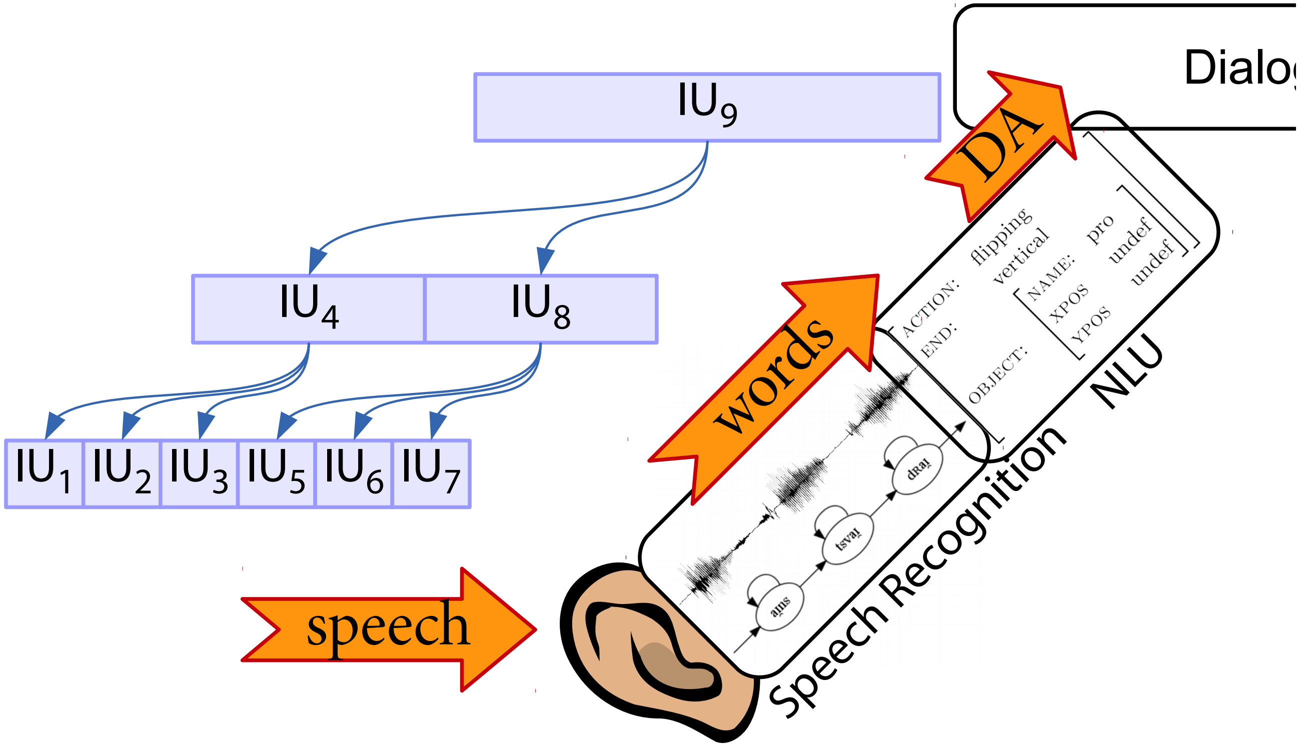
- processing modules are connected via buffers
- buffers contain incremental units (IUs)



  - **grounded-in** links (*grin*) denote ancestry
  - **same-level** links (*sll*) for information of the same type

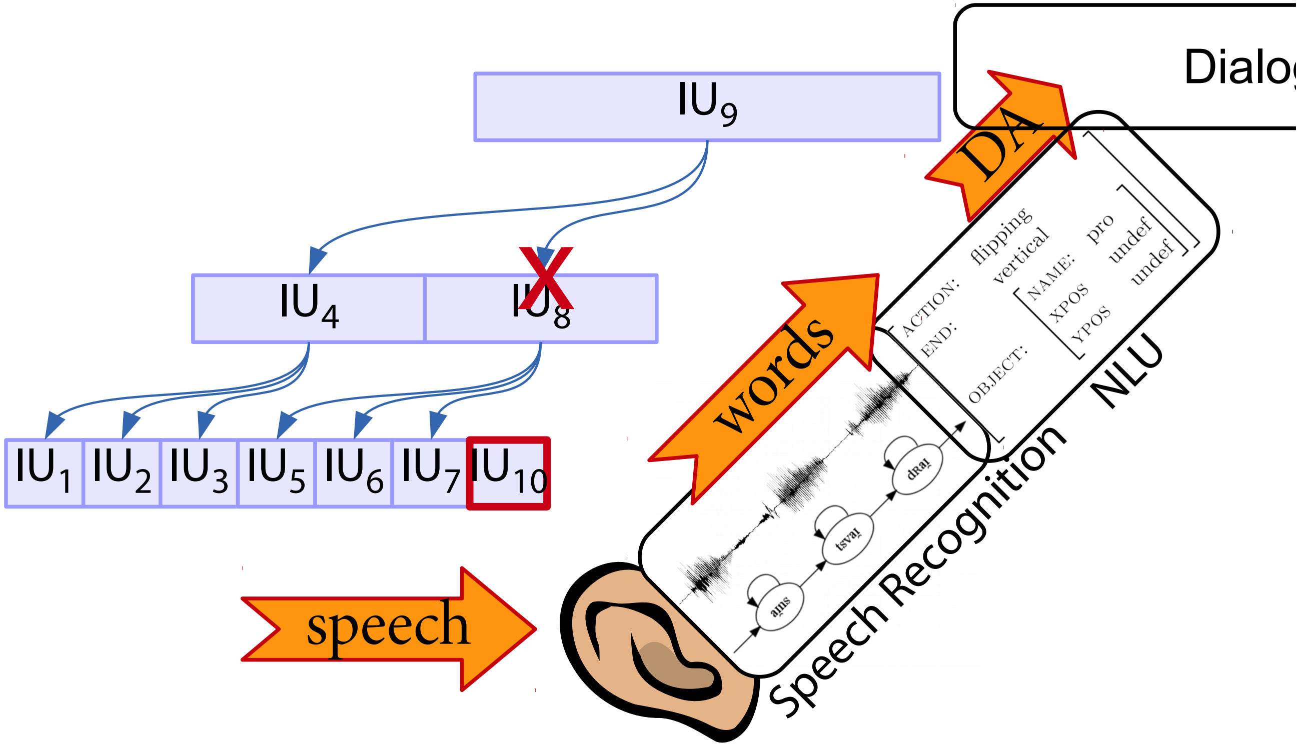
Input Pipeline

- different IU types on different levels
  to denote different kinds of information, e.g.
  - DAs
  - words
  - phonemes



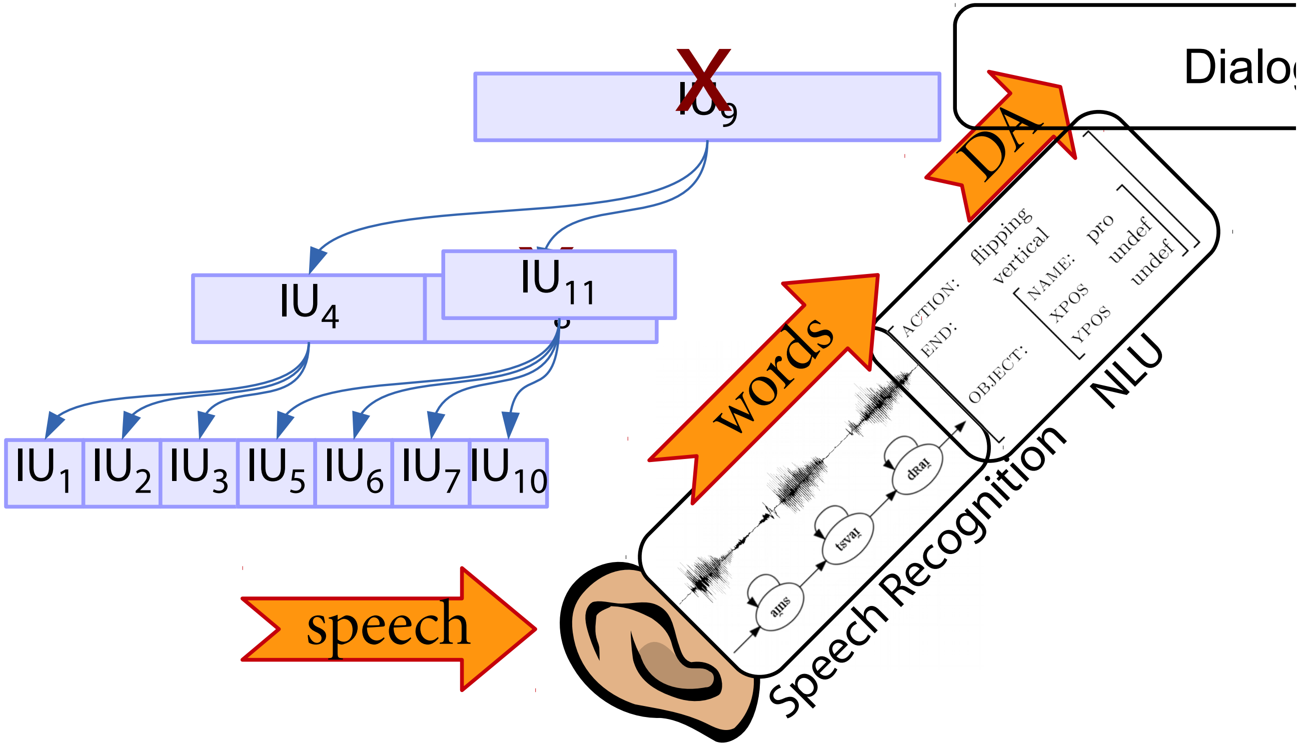
edits as a result of belief changes

- belief changes lead to changes in the network
  - a new frame arrives
  - the word hypothesis is revoked …



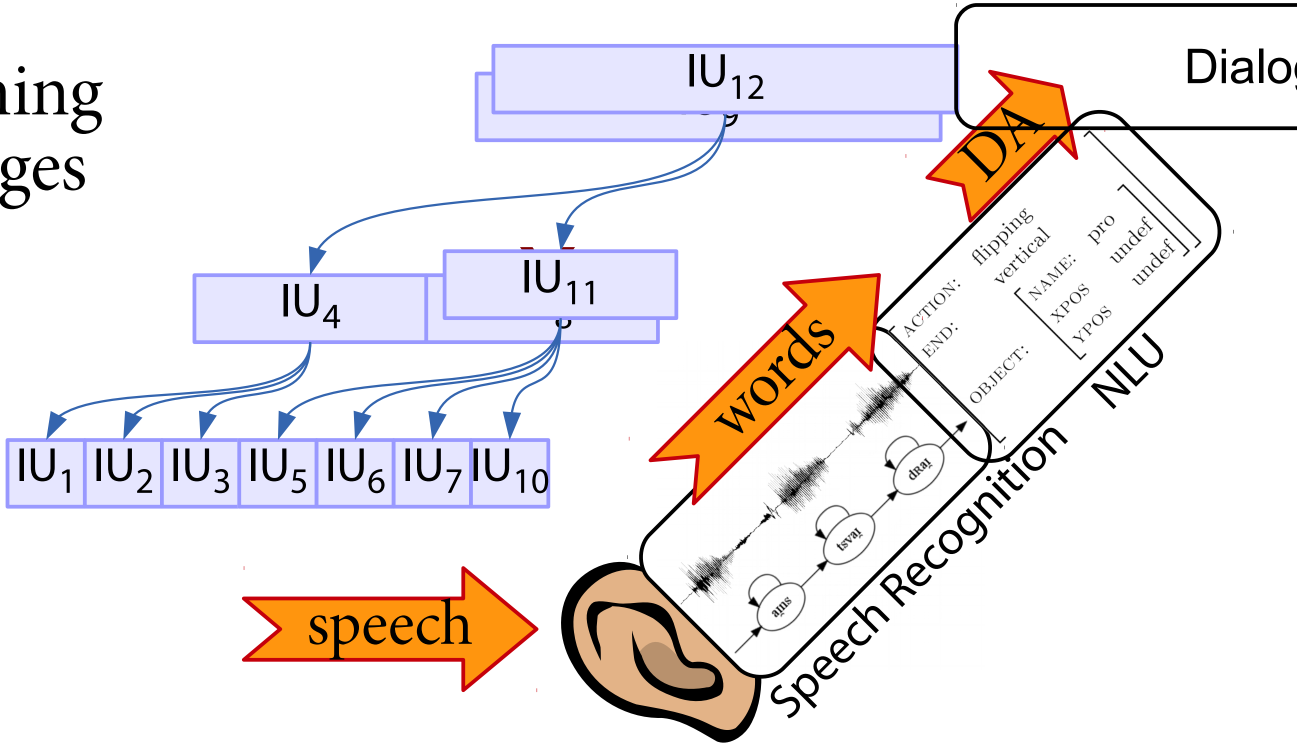
edits as a result of belief changes

- belief changes lead to changes in the network
  - a new frame arrives
  - the word hypothesis is revoked and replaced by a different one



edits as a result of belief changes

- belief changes lead to changes in the network
  - changes trickle up in the system
  - higher-level reasoning might lead to changes trickling down

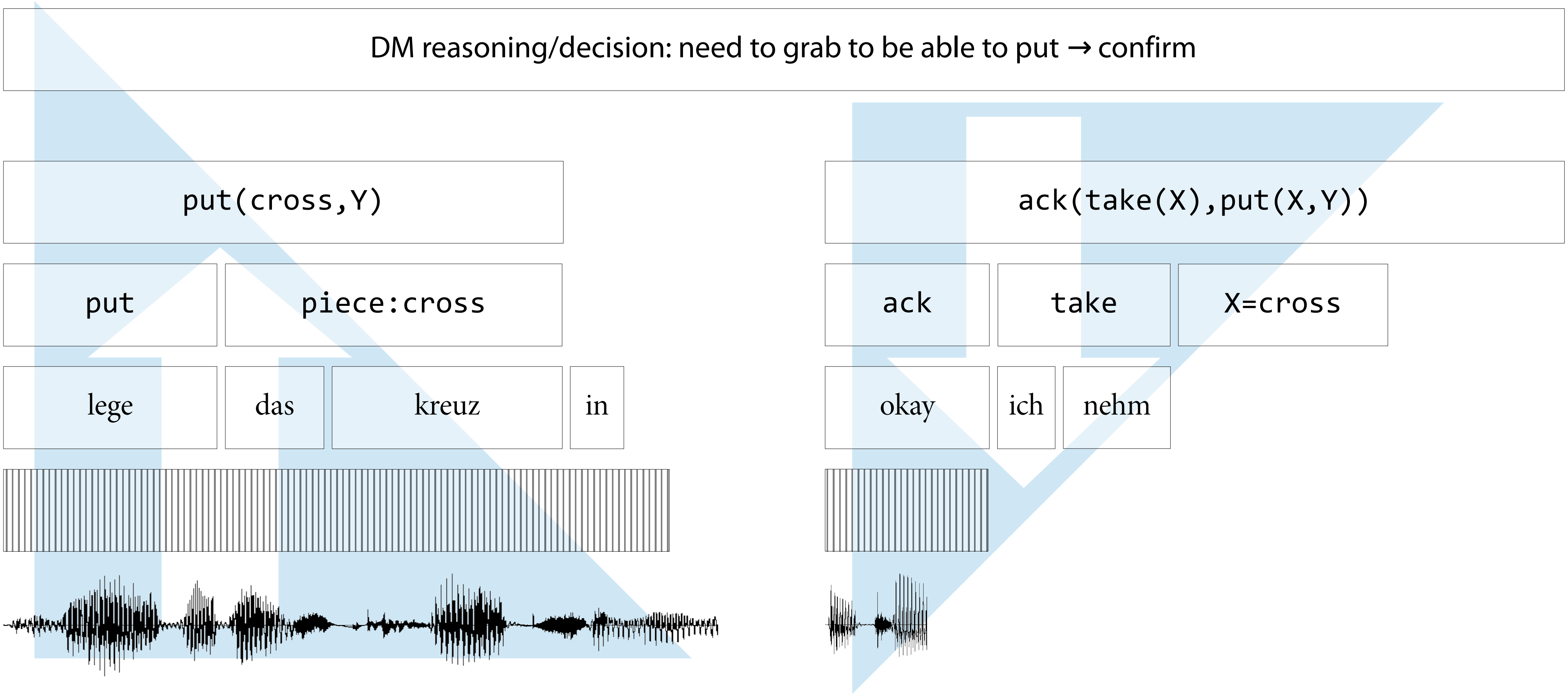


IU Data Model

- Incremental Units (IUs)
  - encapsulate minimal amounts of information
    at the current level of abstraction (phones, words, ideas, …)
  - linked to other units on the *same level* to form hypotheses
  - linked to units they are based on to track dependencies
  - network of units stores information states
- Updates to the network reflect changes in understanding:
  - add units when new information becomes available
  - *revoke* units if they turned out to be wrong
  - notify about degree of commitment/certainty to a unit

Schlangen & Skantze (2009, 2011)

A data model for incremental just-in-time processing



DM reasoning/decision: need to grab to be able to put → confirm

put(cross,Y)

put

piece:cross

lege

das

kreuz

in

ack(take(X),put(X,Y))

ack

take

X=cross

okay

ich

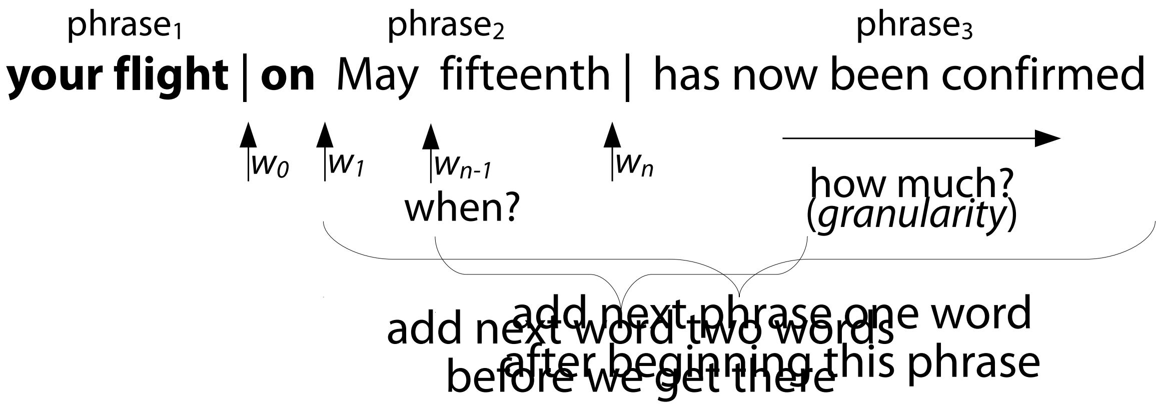
nehm

input side

output side

Incremental Processing: Important Concepts

- **Lookahead:** the amount of context into the future that a processor needs in order to produce (reasonable) output
- **Granularity:** the size of input that is added at a time



- both lower lookahead and finer granularity help to reduce processing delays

The volatility of incremental hypotheses

- incremental hypotheses are often only preliminary
  - four
    fourty
    fourteen
    four teens?
- also long-range dependencies:
  - the horse raced past the barn fell

the	horse	raced	past	the	barn	fell
DT	NN	~~VBD~~	IN	DT	NN	VBD
		VBN				

→ potentially infinite number and span of changes

Example system: incremental input processing

More natural human-computer interaction

- partial incremental (multi-modal) dialogue systems
  - reduced system domains that exploit only one specific aspect
- some example systems
  - subtle feedback to signal understanding, sub-turn interaction
  - the use of affordances in continuous control
  - flexible delivery of spoken output to bind with other modalities
  - flexible spoken output in a noisy domain
  - ability to co-complete / shadow user speech

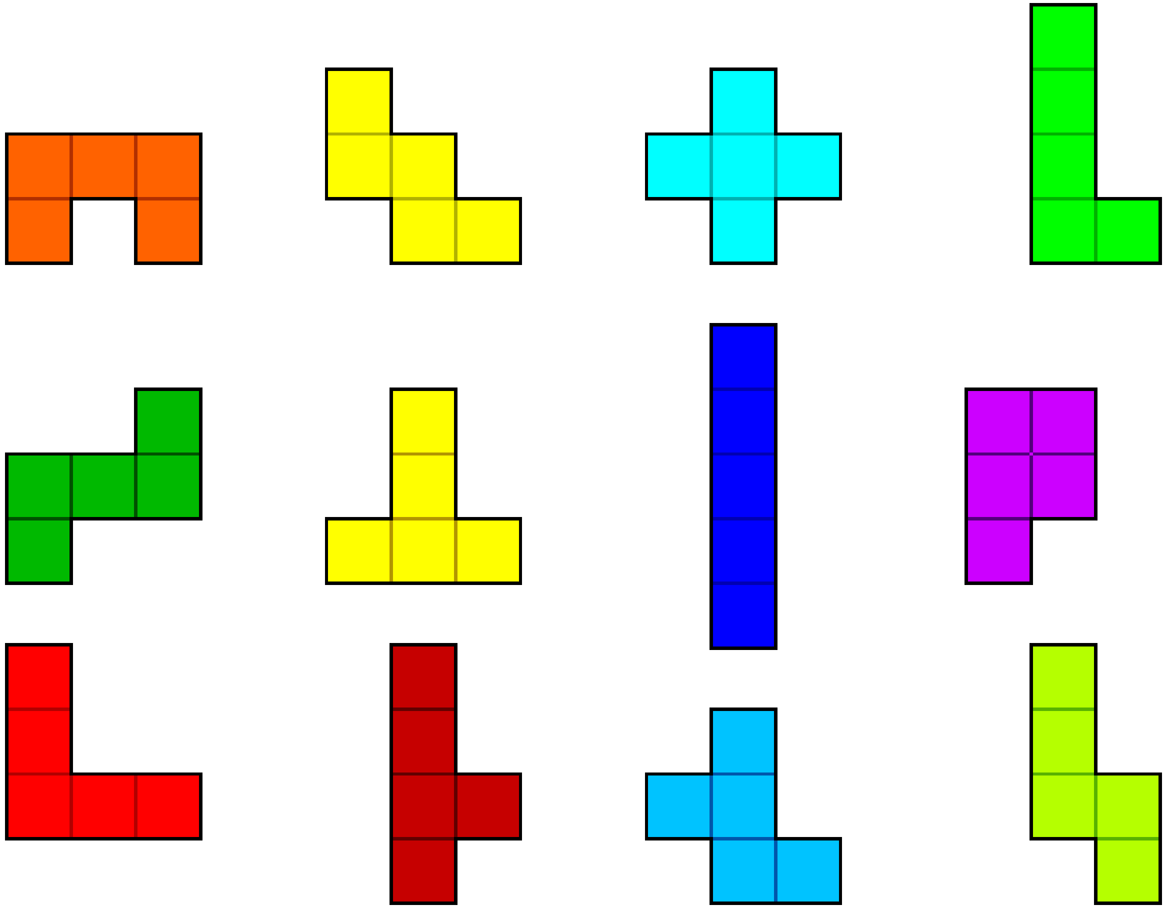
for the „micro-domain principle“ see (Edlund et al, Speech Communication 2008).

Feedback and sub-turn interaction

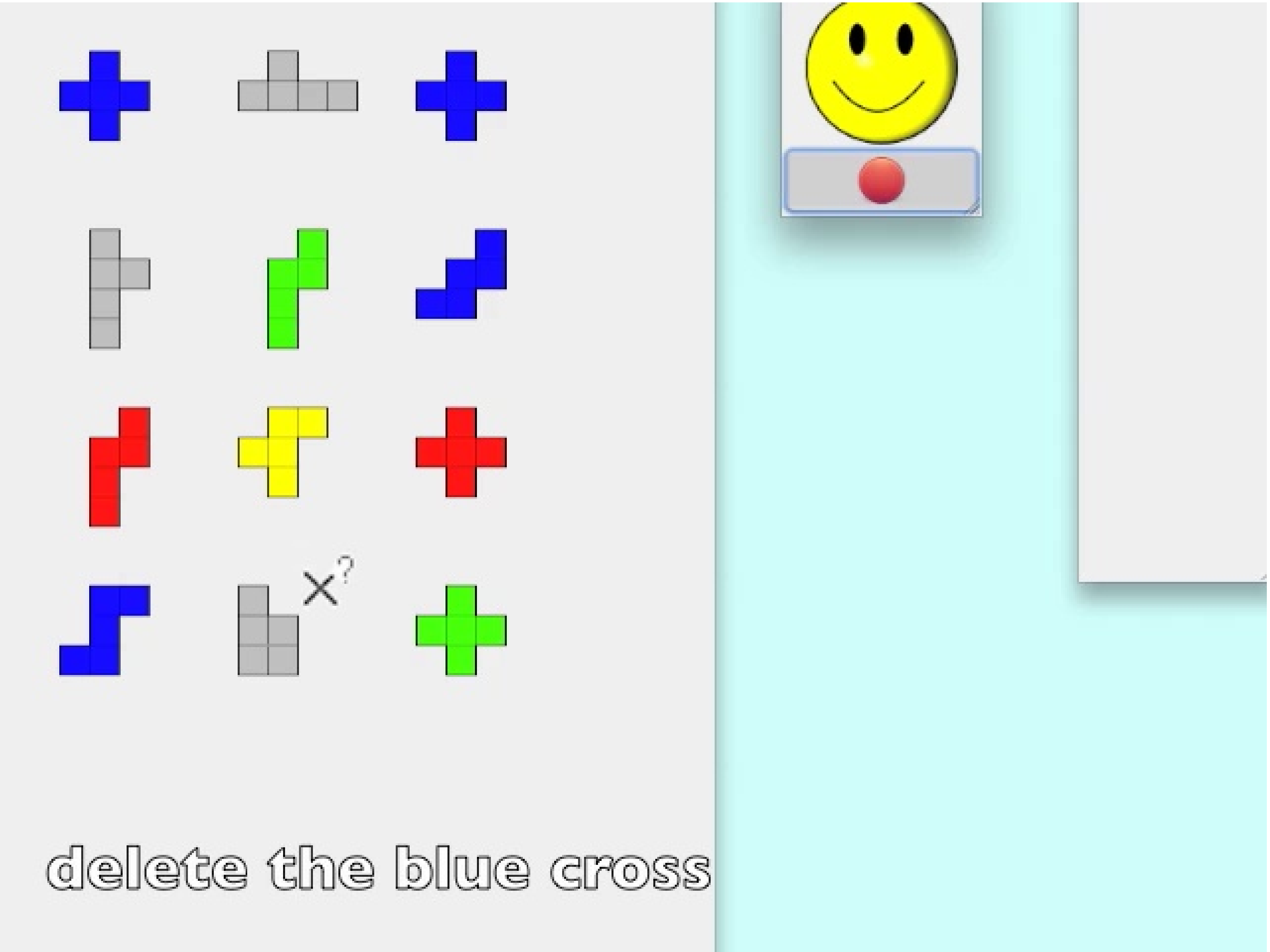
- Humans use feedback to signal state of understanding
  - often within a very tight *feedback loop*
  - incremental processing allows to tighten this feedback loop
  - in the video (to follow): visual feedback during the utterance
- Human reaction time (and type of reaction) depends on pragmatic completeness and prosody
  - crudely modelled using a simple prosodic rule
  - actions are performed as soon as system is certain

A simple task domain

- 12 pentomino pieces
- human is to manipulate pieces:
  - rotate
  - flip
  - delete



Feedback and sub-turn interaction



delete the blue cross

Feedback and sub-turn interaction

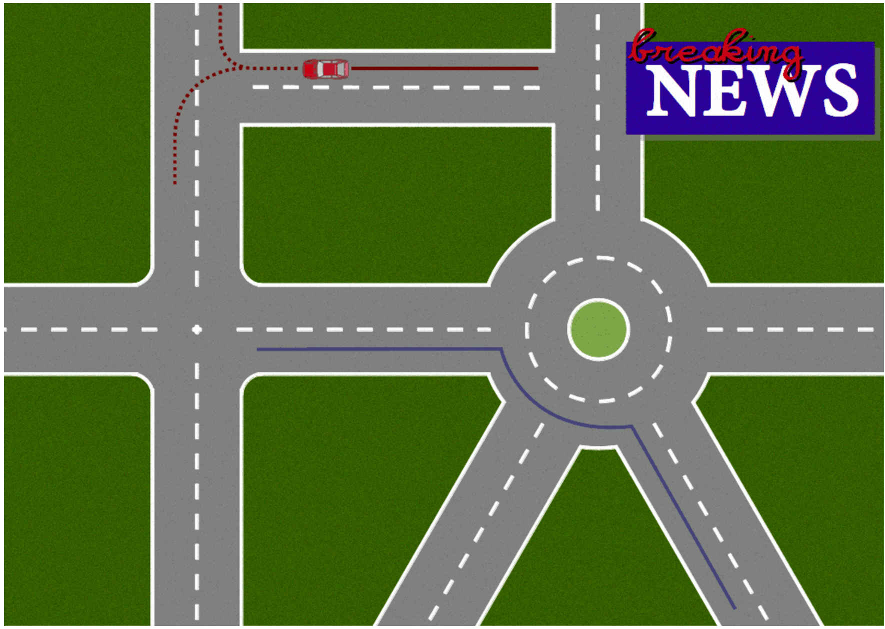
- main features:
  - tight visual feedback loop to signal partial understanding
  - fast, sub-turn interaction based on prosodic rules

- overhearer study showed significantly better rated interactions over a baseline system
  - despite the differences between the systems being very subtle
  - small difference in behaviour → large difference in impression

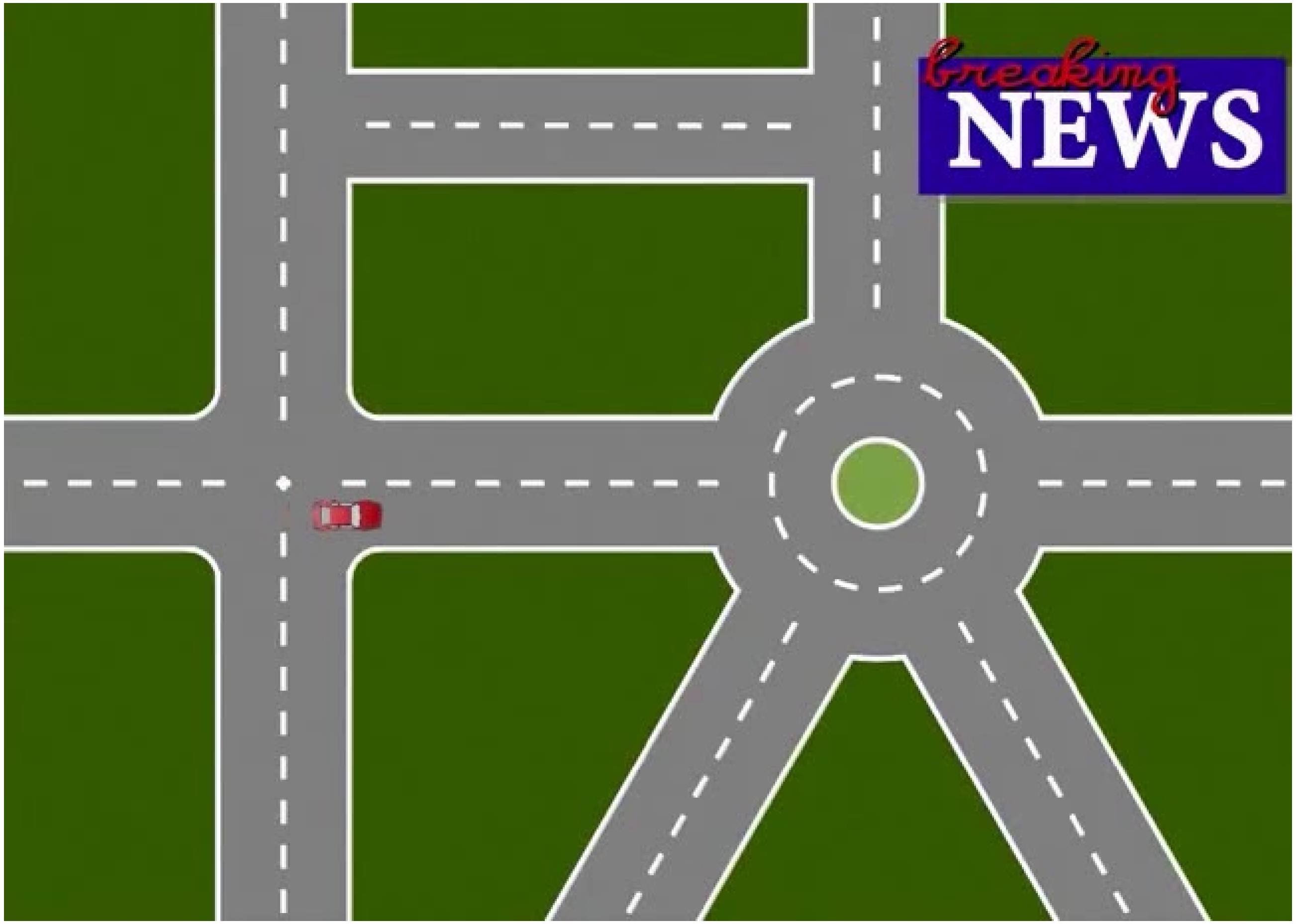
Example system: incremental output processing

Example: The CarChase domain

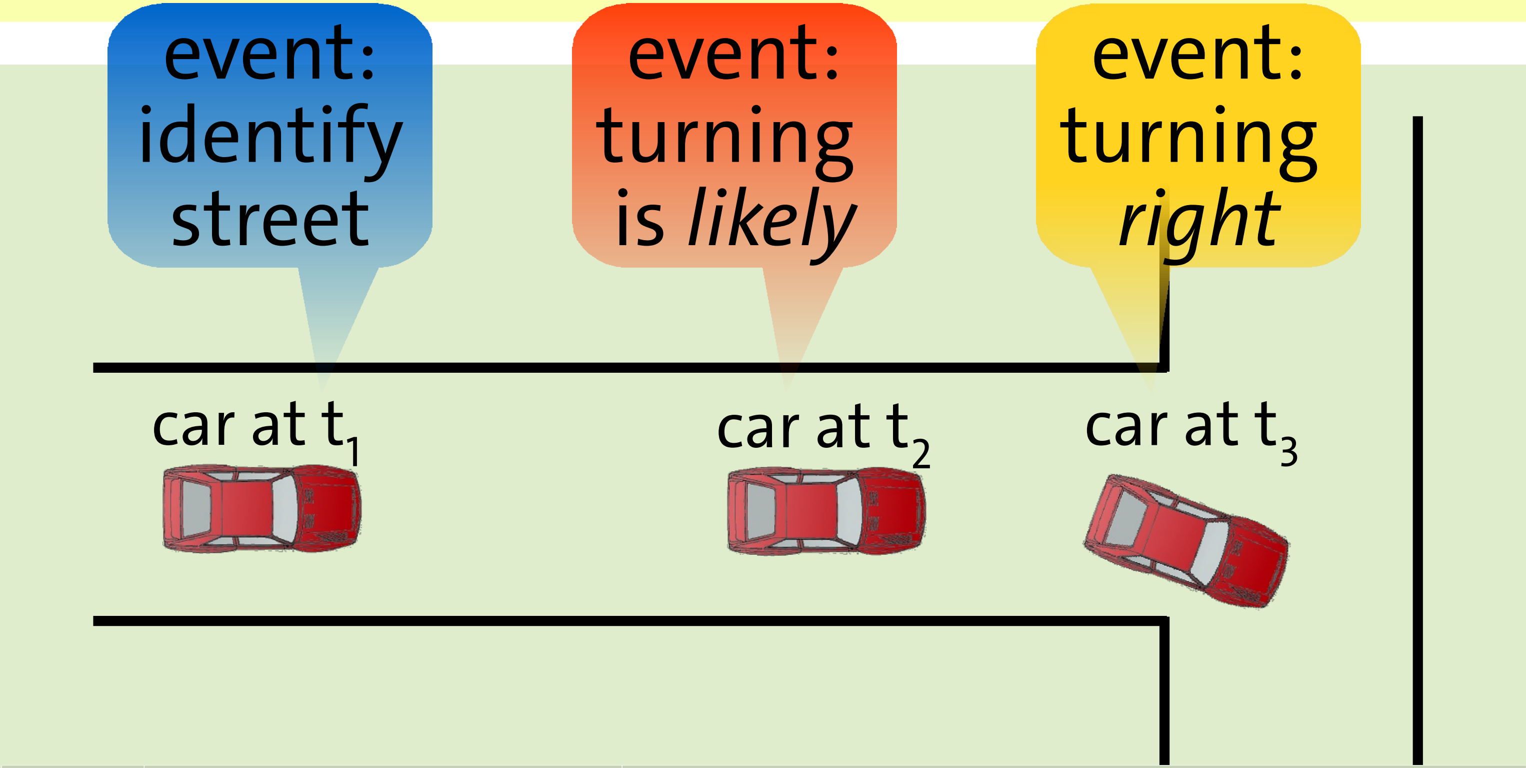
- system comments on events in the scene (car's motion)
- high event rate → impossible to speak isolated utterances
  - combine events into complex utterances (using incremental speech synthesis)
  - skip or abort event notifications in favour of more important information (baseline behaviour)
- simplification of similar real-world scenarios



Standard behaviour



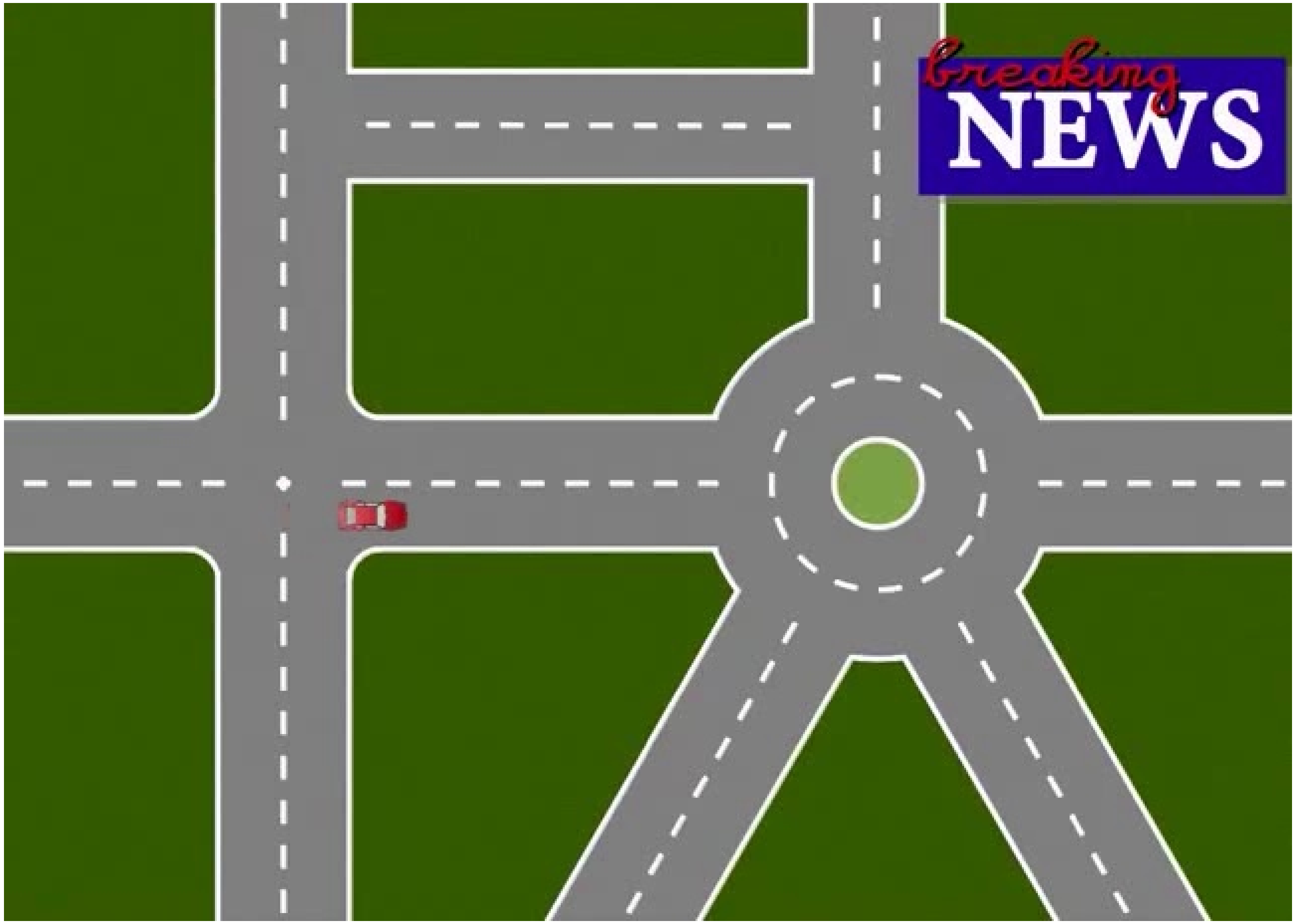
Taking expectations into account



time	event description	ongoing utterance (spoken part in **bold**)
t_1	car on Main Street	**Th**e car drives along Main Street.
t_2	car will have to turn	... **Main Str**eet and then turns ‹hes›
t_3	car turns right	... **Main Street and the**n turns right.

more details on interaction strategy in Baumann&Schlangen, SigDial 2013.

Incremental behaviour (taking expectations into account)



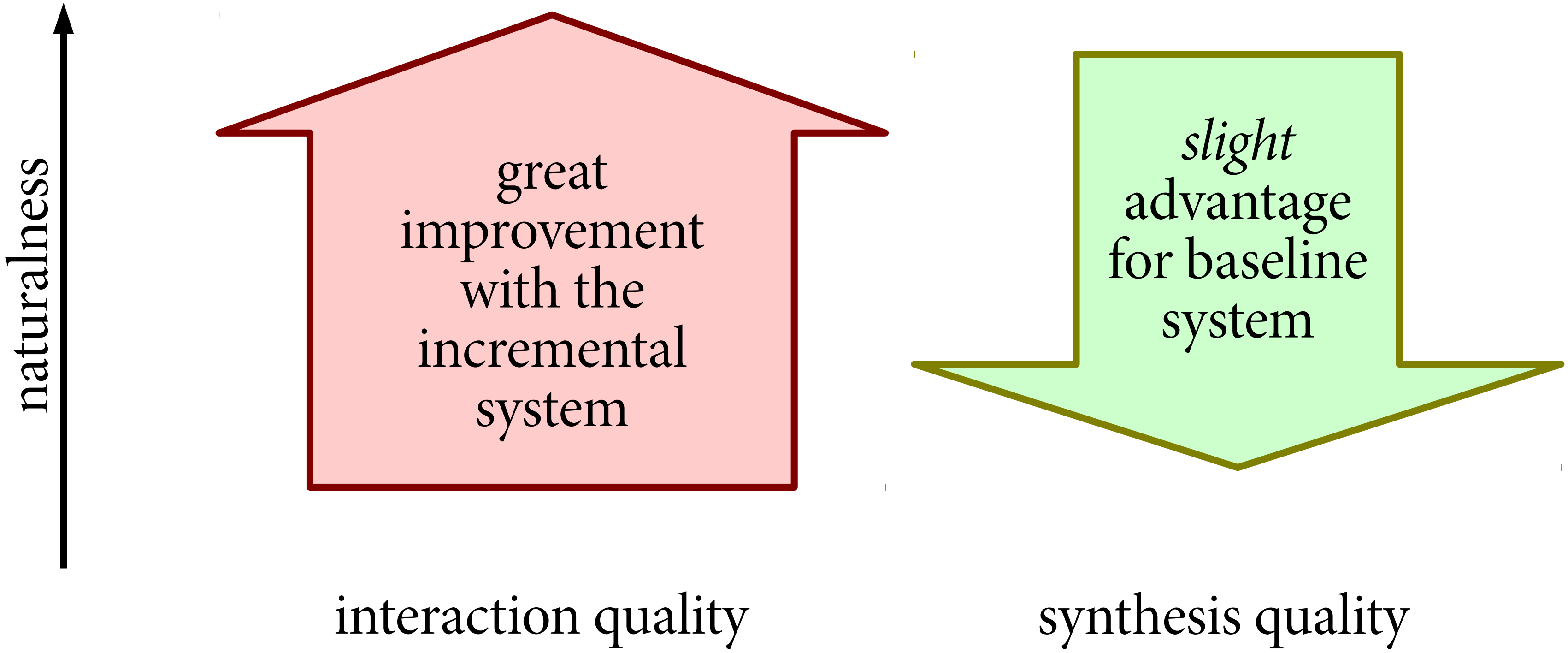
Experiment

- incremental system vs. baseline system
- 9 settings in the CarChase domain
- 9 subjects were asked to rate (5-point Likert)
  - naturalness of verbalization (to capture interactional adequacy)
  - naturalness of *pronunciation* (to capture synthesis quality)
- results in 81 paired samples
- incremental processing implemented in InproTK, using speech synthesis technology from MaryTTS

InproTK: Baumann&Schlangen, SDCTD 2012; MaryTTS: Schröder&Trouvain, IJST 2003.

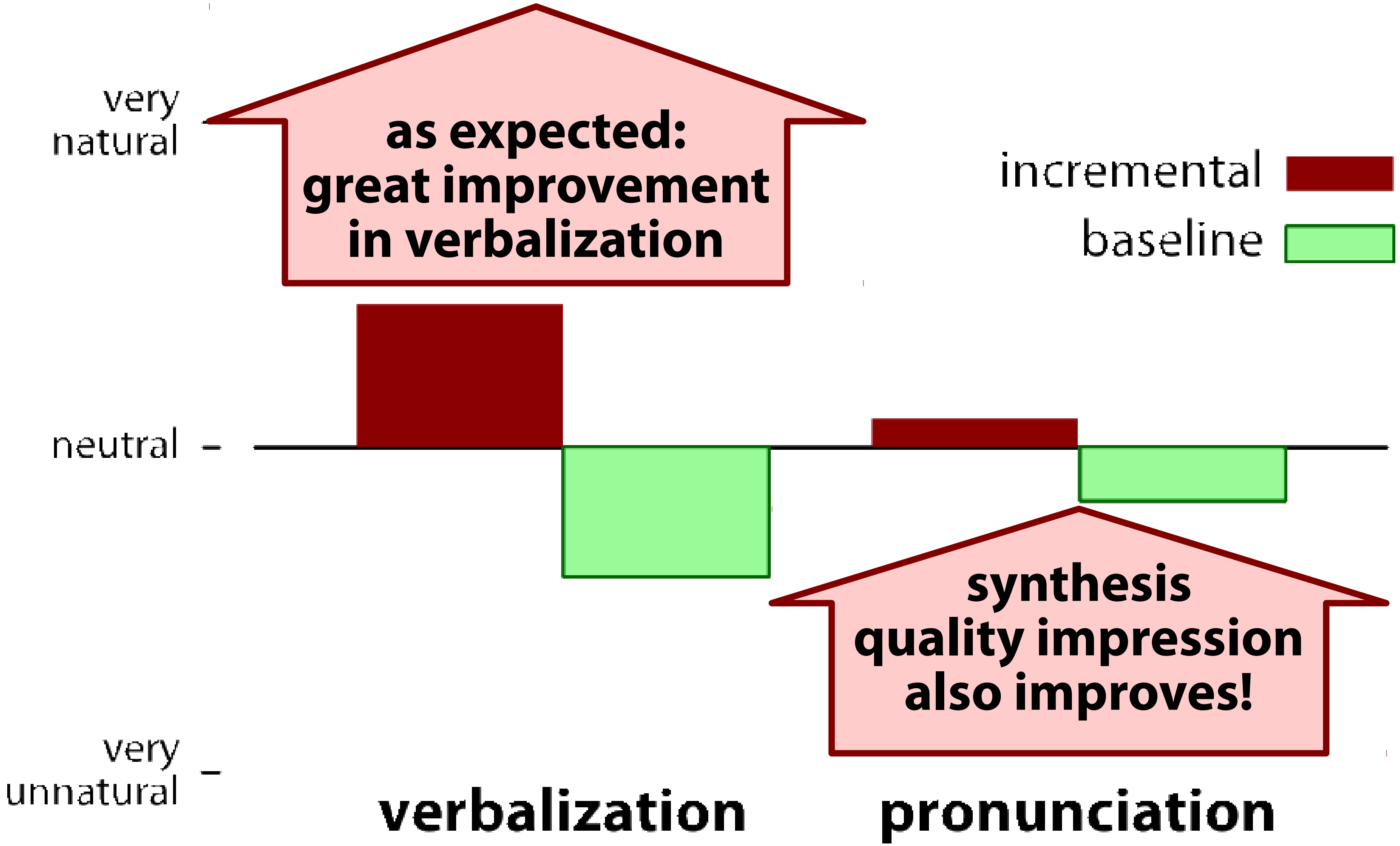
Expected results

- we were hoping for a good trade-off:


naturalness
great improvement with the incremental system
slight advantage for baseline system
interaction quality
synthesis quality

→ write paper: „Trade-off between incrementality of behaviour and speech synthesis quality“

Actual results



as expected: great improvement in verbalization

synthesis quality impression also improves!

Pronunciation ratings

- Incremental processing cannot have systematically improved synthesis quality
  - incremental synthesis was previously shown to lead to a slight quality degradation (Dutoit et al., 2011)
- but:
  naïve listeners do not distinguish between interaction and synthesis quality (Pearson's r = .537)
- verbalization/wording adequacy seems to outweigh pronunciation/synthesis quality

Summary

- processing based on partial input
  - input and output is sub-divided into smaller units
  - output before input is complete
- limited context for decisions (future input missing)
  - allow to *revise* previous hypotheses
- incremental processing enables more natural interaction
  - quick feedback about understanding
  - responsive behaviour

Thank you.

baumann@informatik.uni-hamburg.de

https://nats-www.informatik.uni-hamburg.de/SLP16

Further Reading

- Incremental Processing Architecture:
  - Schlangen, David, and Gabriel Skantze. "A general, abstract model of incremental dialogue processing." Proceedings of EACL, 2009.
- Incremental Speech Recognition, Speech Synthesis, Architecture:
  - Baumann (2013): *Incremental Spoken Dialogue Processing: Architecture and Lower-level Components*. PhD thesis, U Bielefeld, Germany.
- Evaluating Incremental Processing
  - Baumann et al. (2011): "Evaluation and Optimisation of Incremental Processors", *Dialogue & Discourse* **2**(1).
- Highly Interactive Continuous Control
  - Baumann et al. (2013): "Using Affordances to Shape the Interaction in a Hybrid Spoken Dialogue System", *Proceedings of ESSV 2013*, TUD Press.

Notizen

Desired Learning Outcomes

- understand the two dimensions of time involved in incremental processing
- know the incremental unit model and be able to discuss it
- understand the advantage of passing around preliminary information in the system *in a principled way*
- be able to relate incremental processing on various linguistic layers to actual problems in Human-Computer(/Robot) interaction