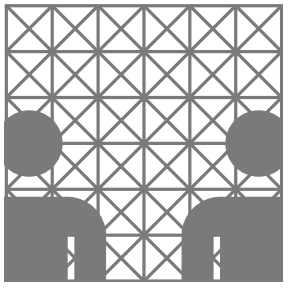
Vorlesung

Sprachdialogsysteme

Timo Baumann
baumann@informatik.uni-hamburg.de

https://nats-www.informatik.uni-hamburg.de/SDS19

Universität Hamburg, Department of Informatics
Language Technology Group



Heute

- Fragen zum Dialogmanagement?
- inkrementelle Sprachverarbeitung

Warum inkrementelle Sprachverarbeitung?

Mensch-Computer Interaktion

Mensch-Computer Interaktion

Navigationsgerät:

„Biegen Sie rechts in 60 Metern ab.“

Mensch-Computer Interaktion

Mensch-Computer Interaktion

Mensch-Computer Interaktion

Gesprochene Sprache

Gesprochene Sprache

Menschen sprechen *responsiv*.

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts musst du abbiegen.“

Menschen sprechen *responsiv*.

1. internes Umplanen

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts ...

Menschen sprechen *responsiv*.

1. internes Umplanen

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts … gradeaus quasi.“

also die zweite.“

Menschen sprechen *responsiv*.

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts ...

Menschen sprechen *responsiv*.

2. externe Ereignisse

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts ...

Menschen sprechen *responsiv*.

2. externe Ereignisse

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts … dem blauen Golf hinterher.“

Menschen sprechen *responsiv*.

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts ...

Menschen sprechen *responsiv*.

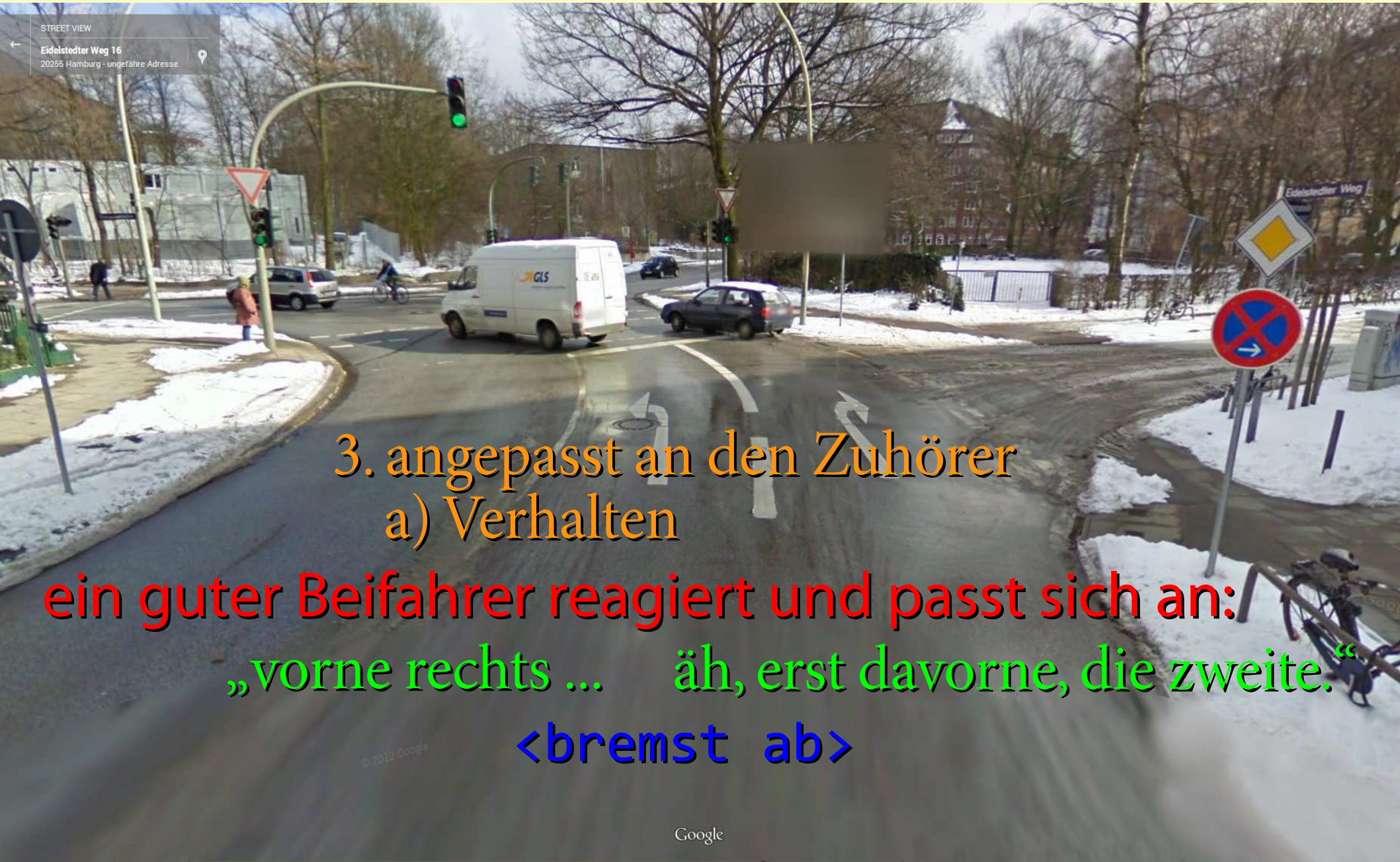
3. angepasst an den Zuhörer

a) Verhalten

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts ...

Menschen sprechen *responsiv*.



3. angepasst an den Zuhörer
a) Verhalten

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts ... äh, erst davorne, die zweite.“

<bremst ab>

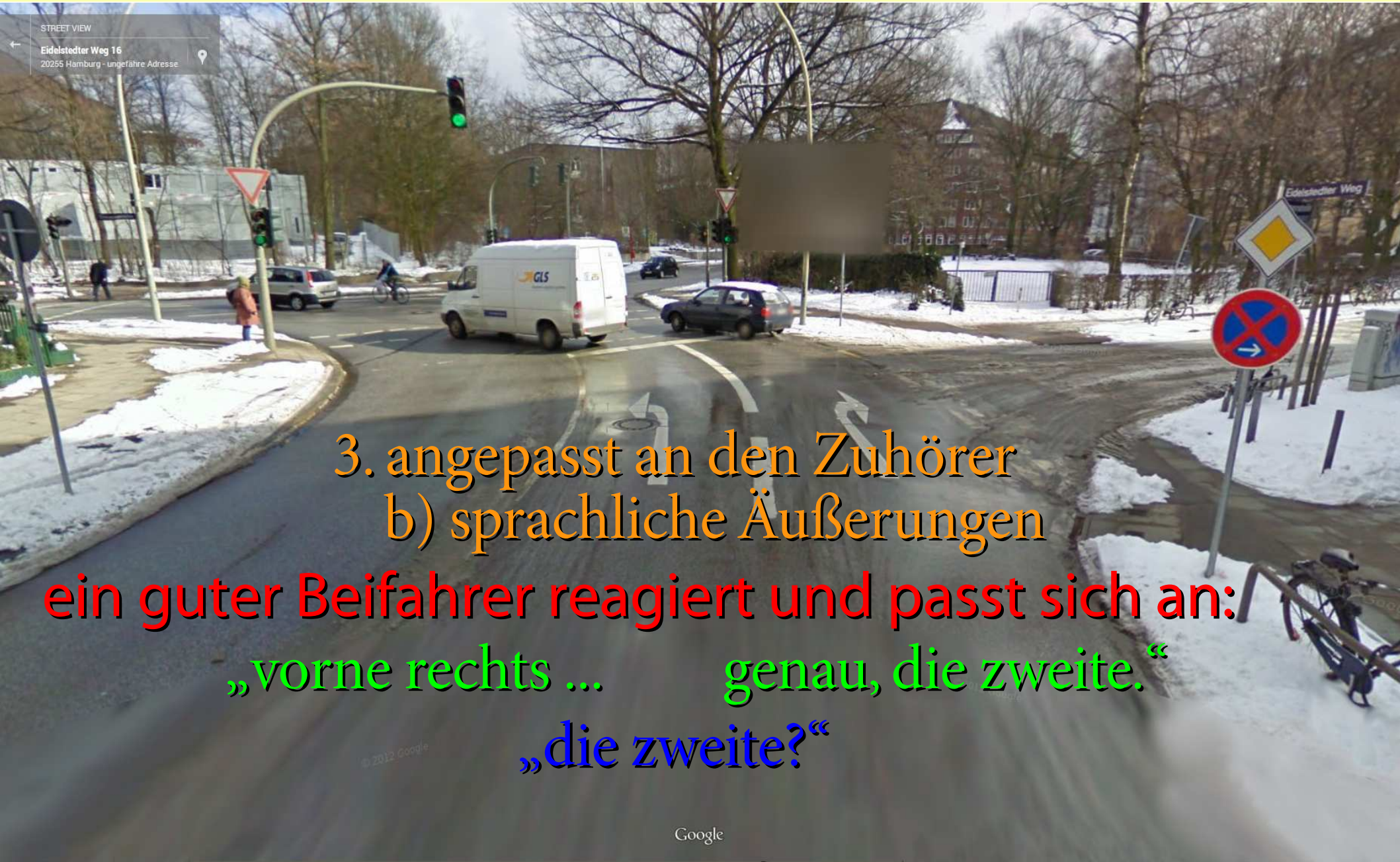
Menschen sprechen *responsiv*.

3. angepasst an den Zuhörer
b) sprachliche Äußerungen

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts ...

Menschen sprechen *responsiv*.



3. angepasst an den Zuhörer
b) sprachliche Äußerungen

ein guter Beifahrer reagiert und passt sich an:

„vorne rechts ... genau, die zweite.“

„die zweite?“

Zwischenfazit zur Sprachinteraktion

* bzw. nur sehr eingeschränkt (Baumann 2013 Diss, Baumann&Schlangen 2013, ...)

Zwischenfazit zur Sprachinteraktion

1. Sprache entwickelt sich über die Zeit
2. die Welt ändert sich, während wir sprechen
3. wir ändern uns, während mit uns gesprochen wird

   → Fähigkeit während des Sprechens umzuplanen

   → Fähigkeit während des Zuhörens zu planen & agieren

- Lösung: schritthaltende (inkrementelle) Verarbeitung ermöglicht Responsivität in einer veränderlichen Welt

➔ Computersysteme sind dazu bisher nicht* fähig

* bzw. nur sehr eingeschränkt (Baumann 2013 Diss, Baumann&Schlangen 2013, ...)

Zwischenfazit zur Sprachinteraktion

1. Sprache entwickelt sich über die Zeit
2. die Welt ändert sich, während wir sprechen
3. wir ändern uns, während mit uns gesprochen wird

   → Fähigkeit während des Sprechens umzuplanen

   → Fähigkeit während des Zuhörens zu planen & agieren

- Lösung: schritthaltende (inkrementelle) Verarbeitung ermöglicht Responsivität in einer veränderlichen Welt

➔ Computersysteme sind dazu bisher nicht* fähig

* bzw. nur sehr eingeschränkt (Baumann 2013 Diss, Baumann&Schlangen 2013, ...)

weitere Szenarien

- maschinelles Simultandolmetschen

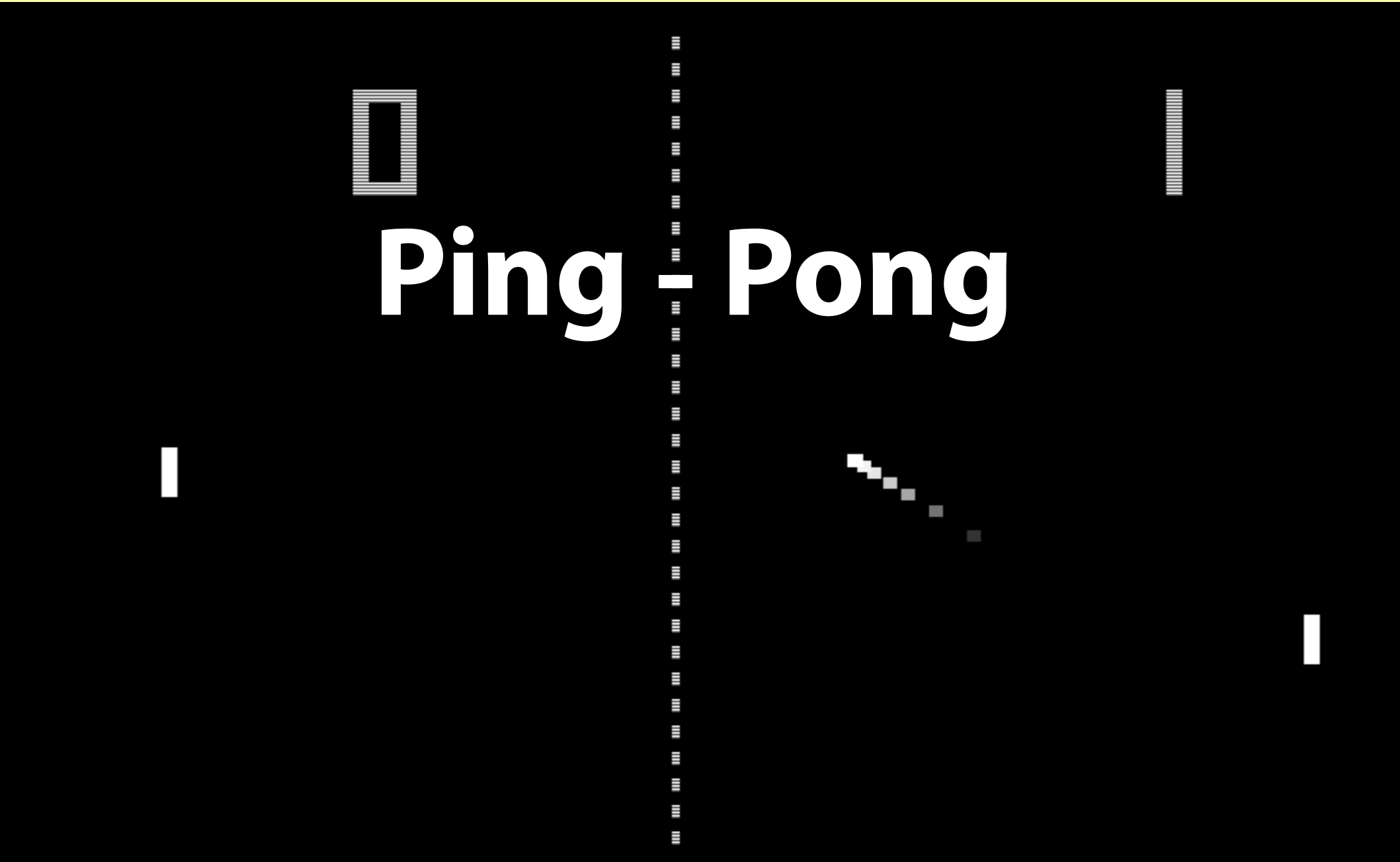
  v.a. internes Umplanen (Baumann et al., 2014 IWSLT)
- Dialoginteraktion Mensch ↔ Roboter

  v.a. externe Ereignisse (Baumann&Lindner, 2015 ICSR)
- Interaktion mit konversationalen Dialogsystemen
  - v.a. Anpassung an Nutzerfeedback
    (Buschmeier et al., 2012, SigDial; Baumann et al., 2013 ESSV)

➔ praktisch jegliche Sprach*interaktion* profitiert von responsivem Verhalten

vorherrschende Form der Mensch-Maschine-Sprachinteraktion

vorherrschende Form der Mensch-Maschine-Sprachinteraktion



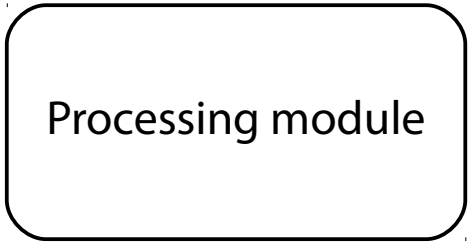
schritthaltende / inkrementelle Verarbeitung: Definition

eine inkrementelle Verarbeitungskomponente

- verarbeitet Eingaben Stück-für-Stück
- generiert (vorläufige) Ausgaben bevor die Eingabeverarbeitung abgeschlossen ist

Incremental vs. Non-incremental Processing

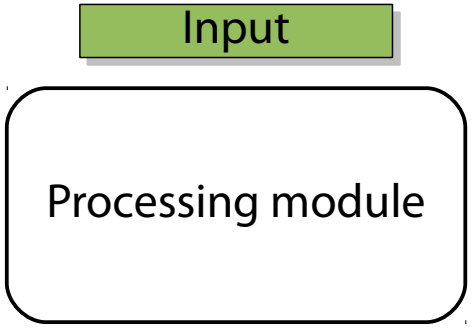
- herkömmliche Verarbeitung:



- Verarbeitung beginnt mit Abschluss der Eingabe → Delay!
  - modulares System: Delays summieren sich

Incremental vs. Non-incremental Processing

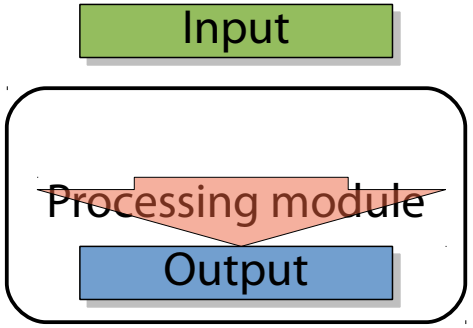
- herkömmliche Verarbeitung:



- Verarbeitung beginnt mit Abschluss der Eingabe → Delay!
  - modulares System: Delays summieren sich

Incremental vs. Non-incremental Processing

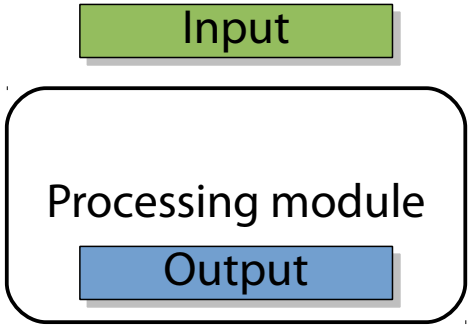
- herkömmliche Verarbeitung:



- Verarbeitung beginnt mit Abschluss der Eingabe → Delay!
  - modulares System: Delays summieren sich

Incremental vs. Non-incremental Processing

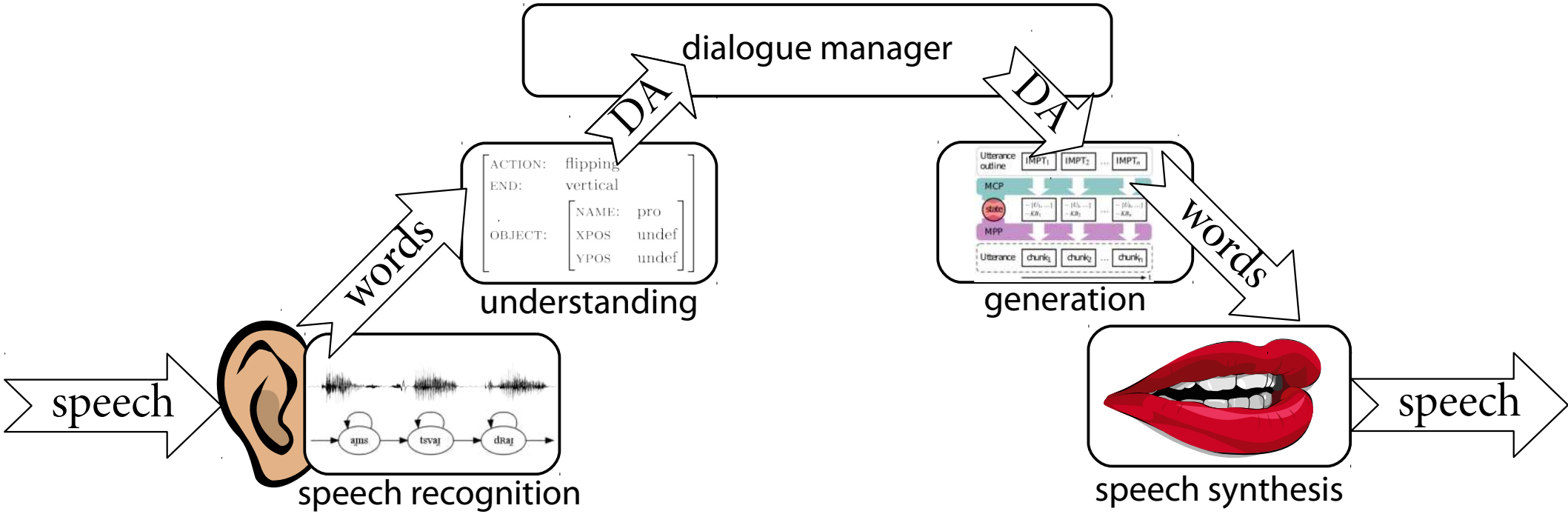
- herkömmliche Verarbeitung:



- Verarbeitung beginnt mit Abschluss der Eingabe → Delay!
  - modulares System: Delays summieren sich

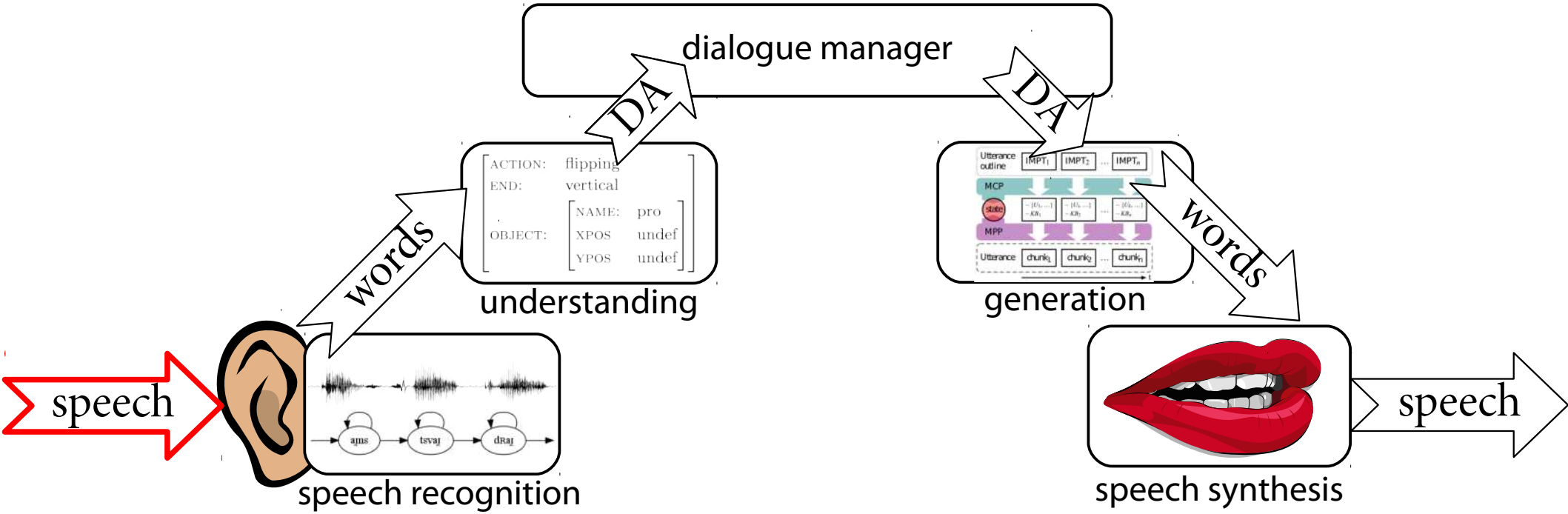
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



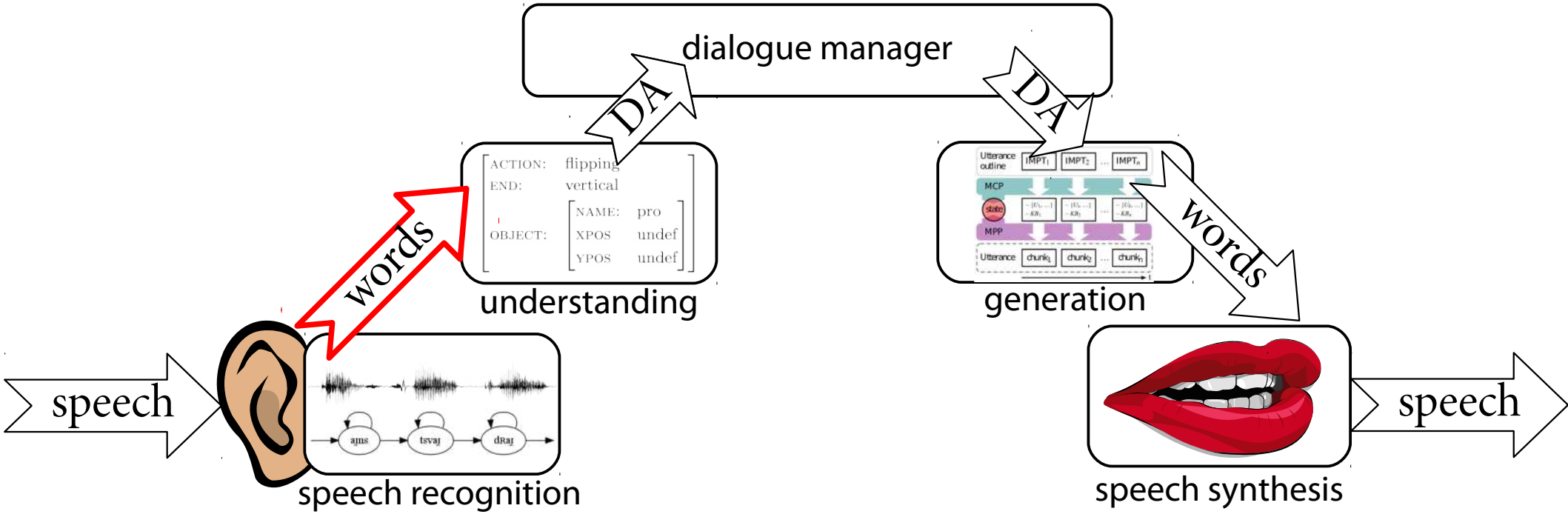
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



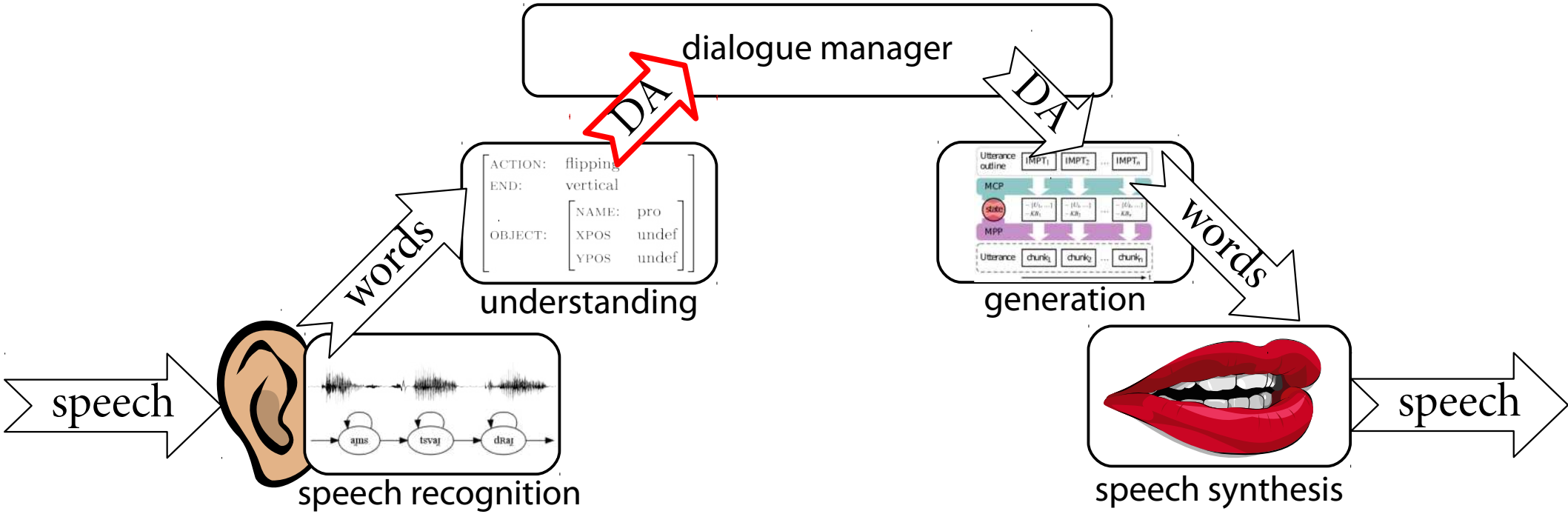
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



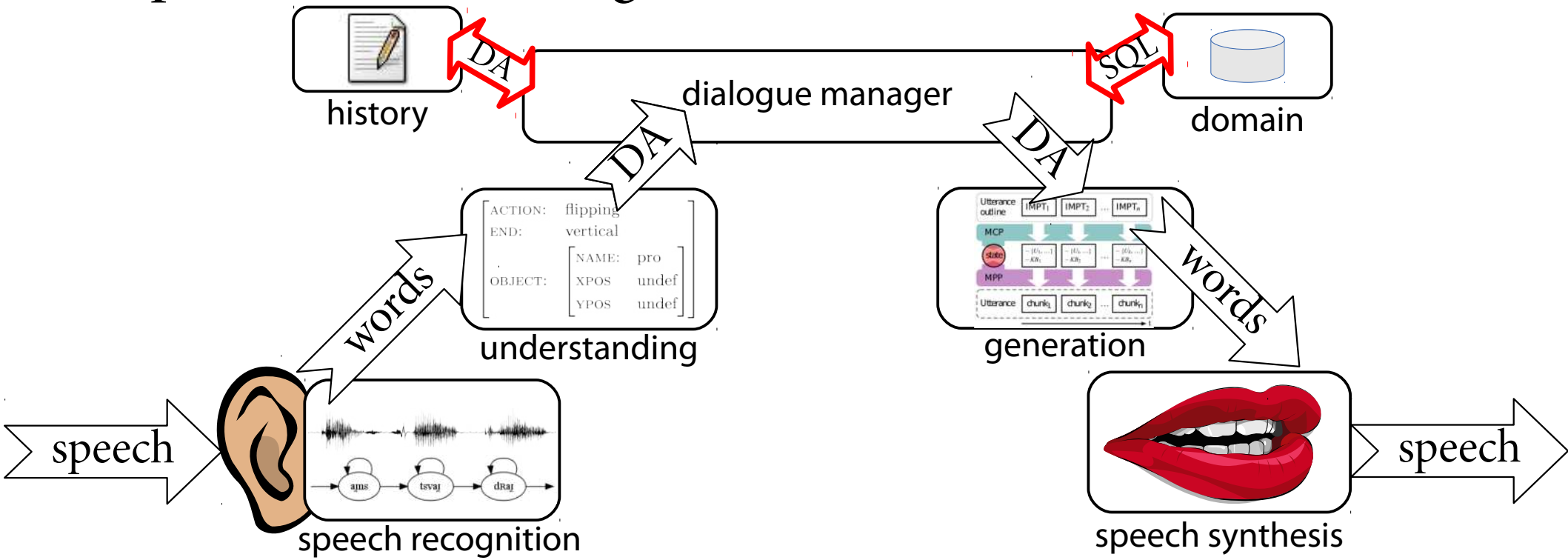
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



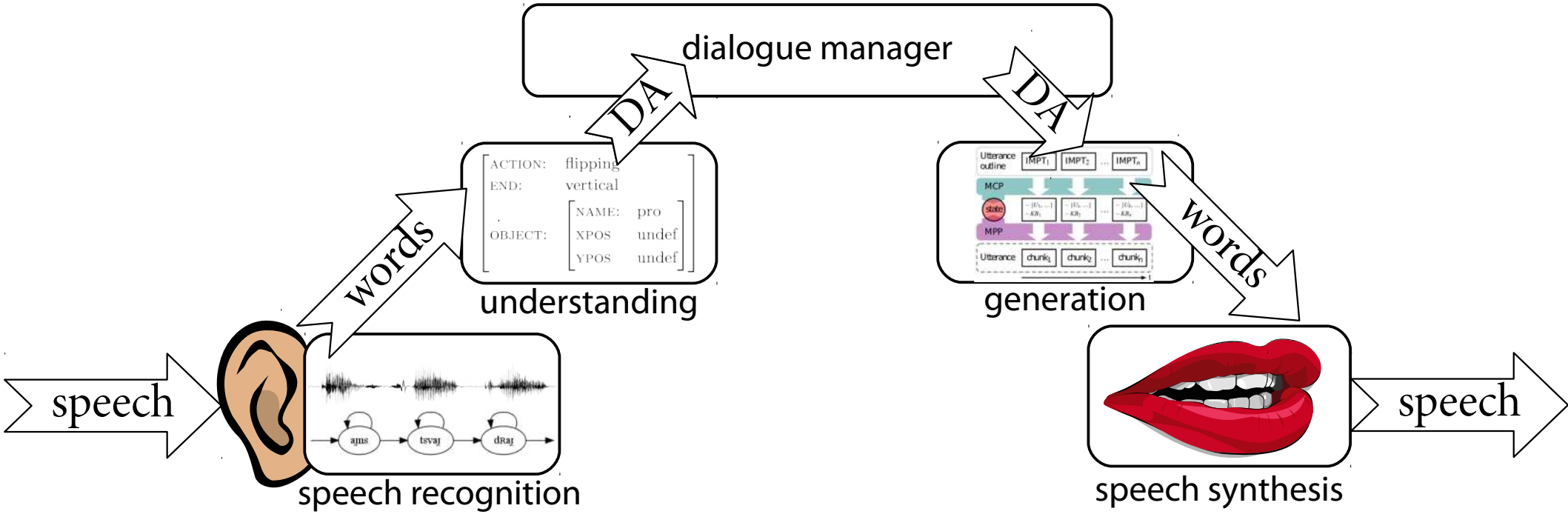
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



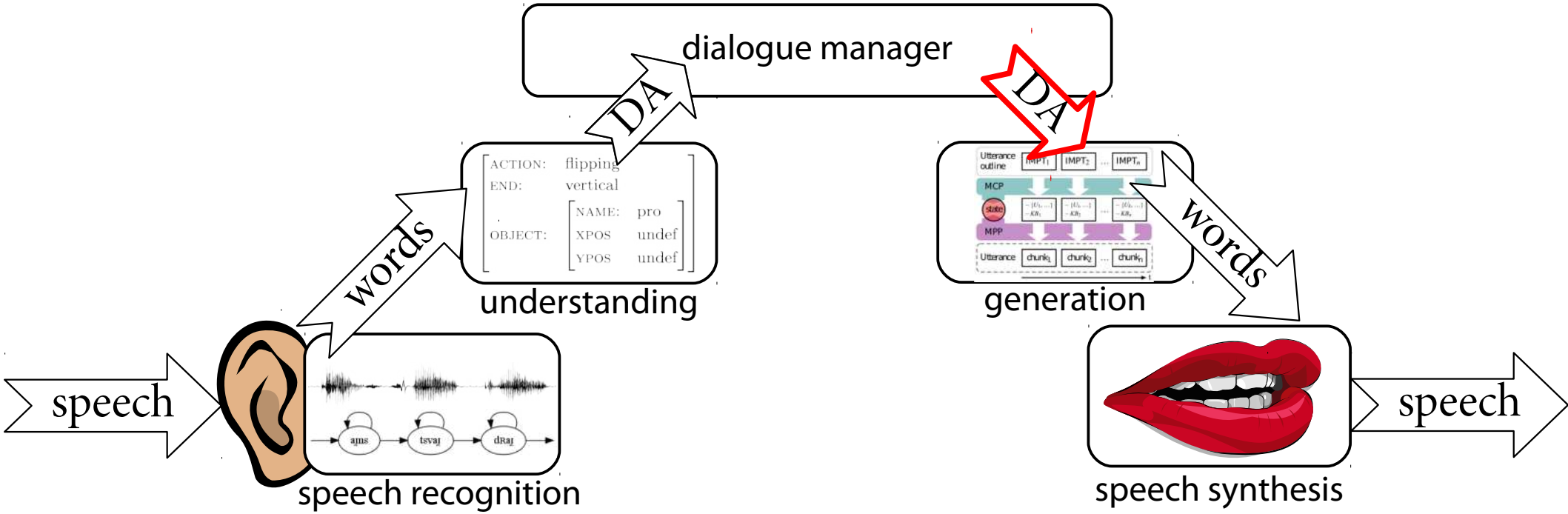
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



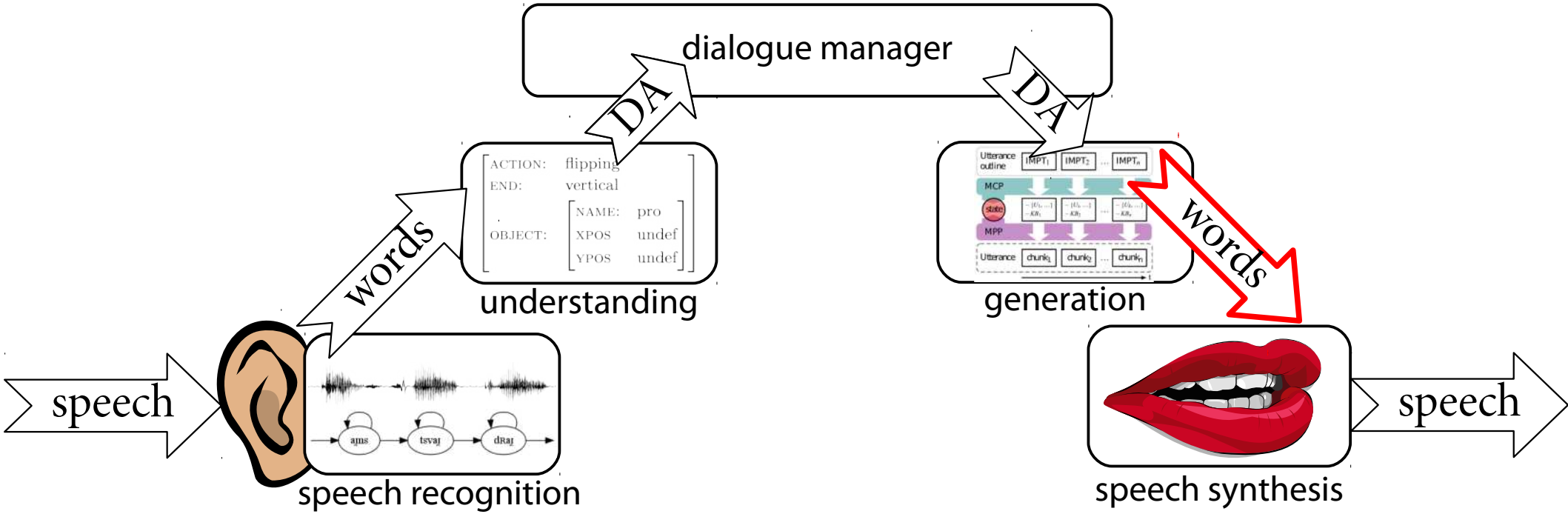
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



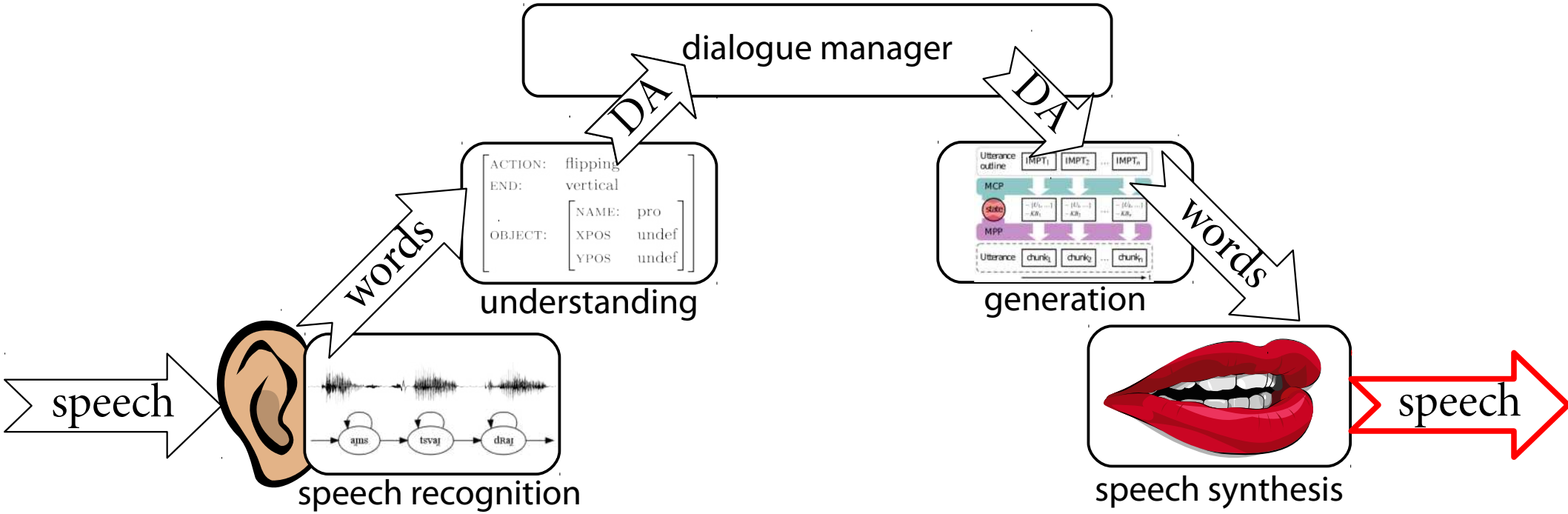
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



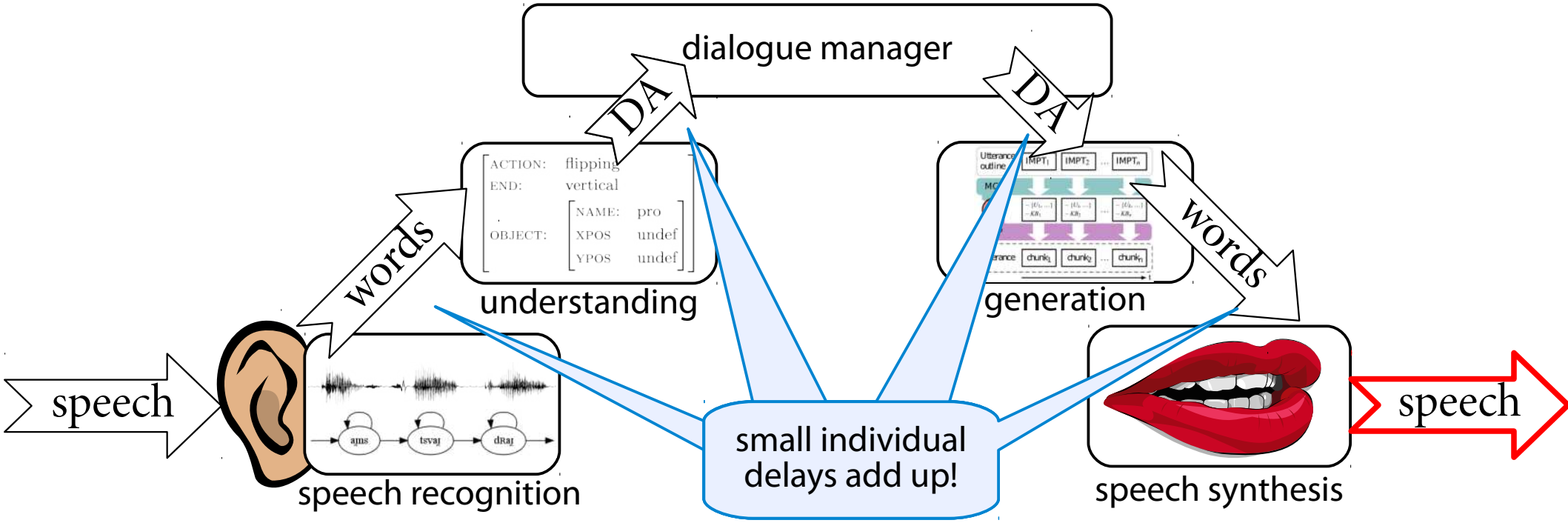
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



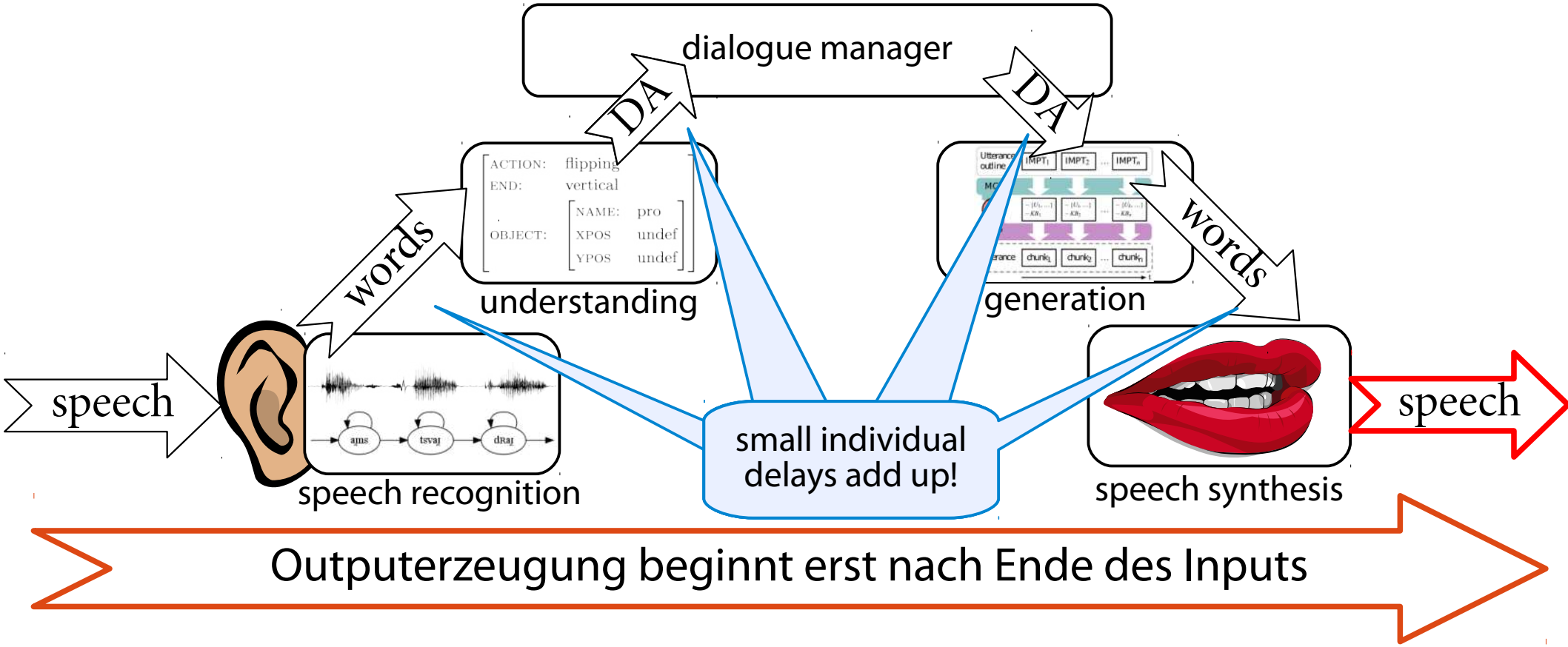
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



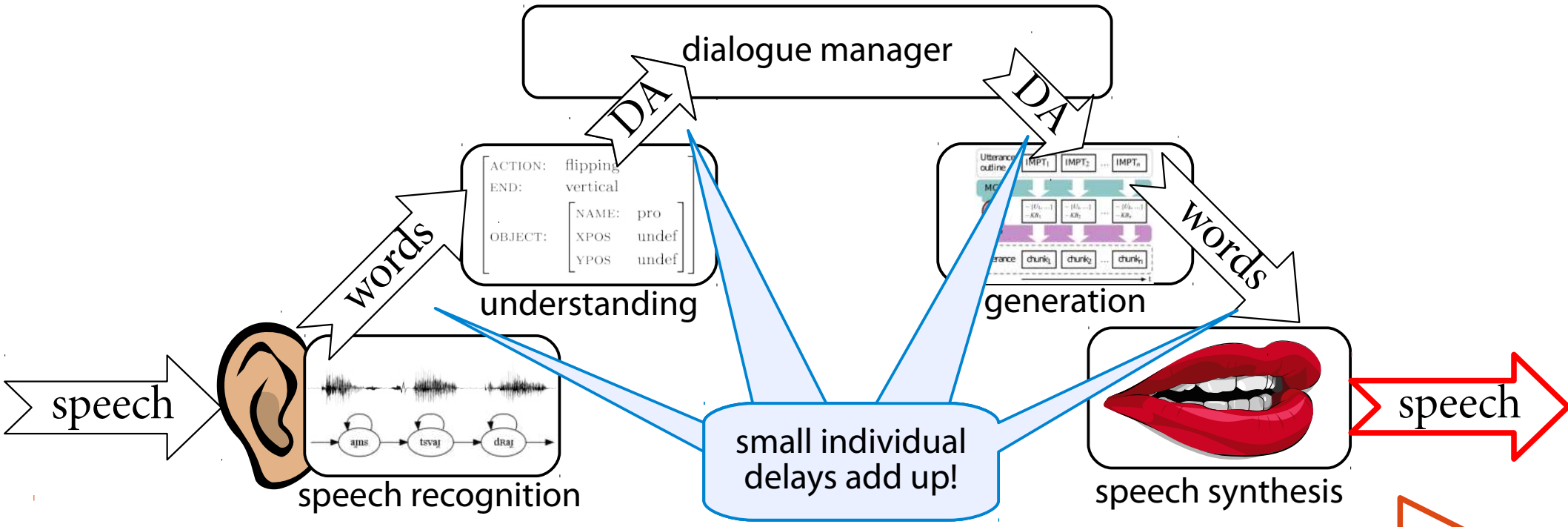
Paradigma: Modulare Verarbeitung

- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



Paradigma: Modulare Verarbeitung

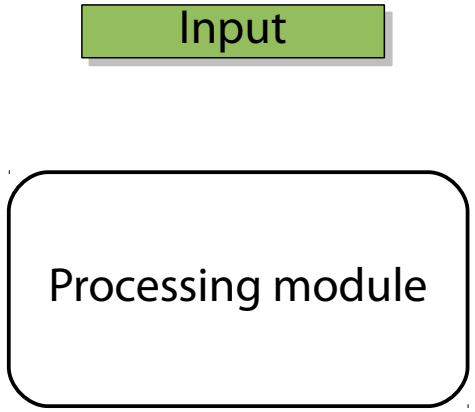
- Verarbeitung gesprochener Sprache ist komplex
  → Teile-und-hersche durch spezialisierte Module
- Pipeline-Verarbeitung, ein Modul nach dem anderen



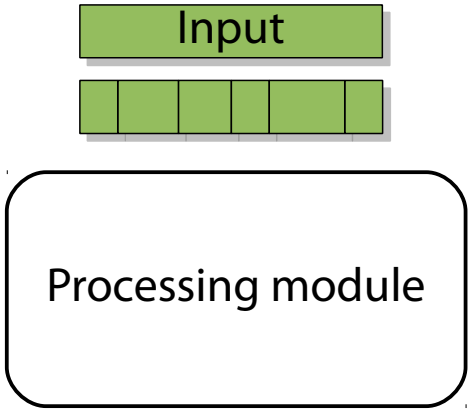
Outputerzeugung beginnt erst nach Ende des Inputs

keine Aktionen vor Ende der Inputverarbeitung

Incremental Processing

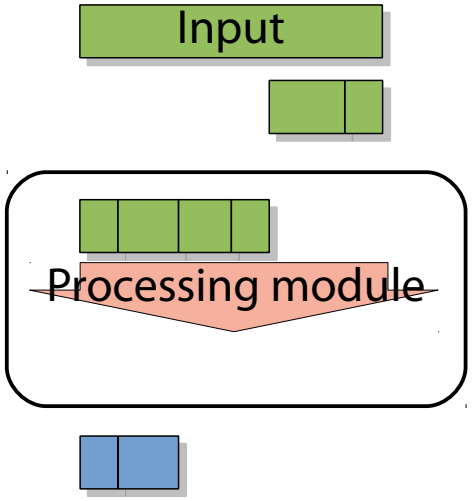


Incremental Processing



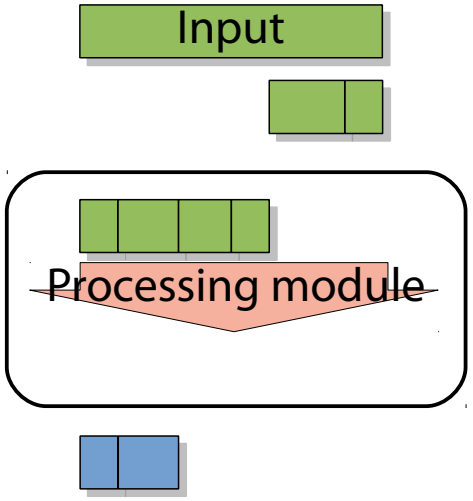
- Eingabe besteht aus minimalen Einheiten die stück-für-stück verarbeitet werden (Sprachaudio, Wörter, Gedanken, …)
- Ausgaben werden schon für Teileingaben erzeugt
- Eingabeeinheiten können zu größeren Ausgabeeinheiten zusammengefasst werden

Incremental Processing



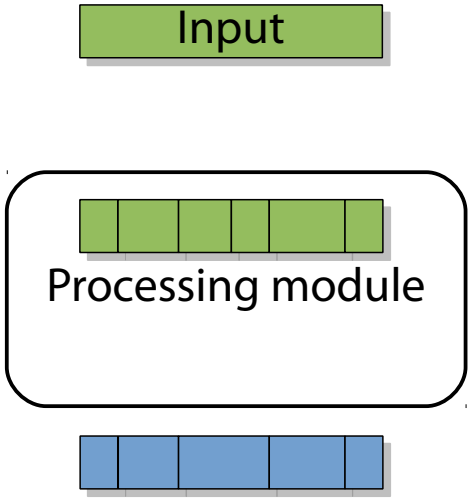
- Eingabe besteht aus minimalen Einheiten die stück-für-stück verarbeitet werden (Sprachaudio, Wörter, Gedanken, …)
- Ausgaben werden schon für Teileingaben erzeugt
- Eingabeeinheiten können zu größeren Ausgabeeinheiten zusammengefasst werden

Incremental Processing



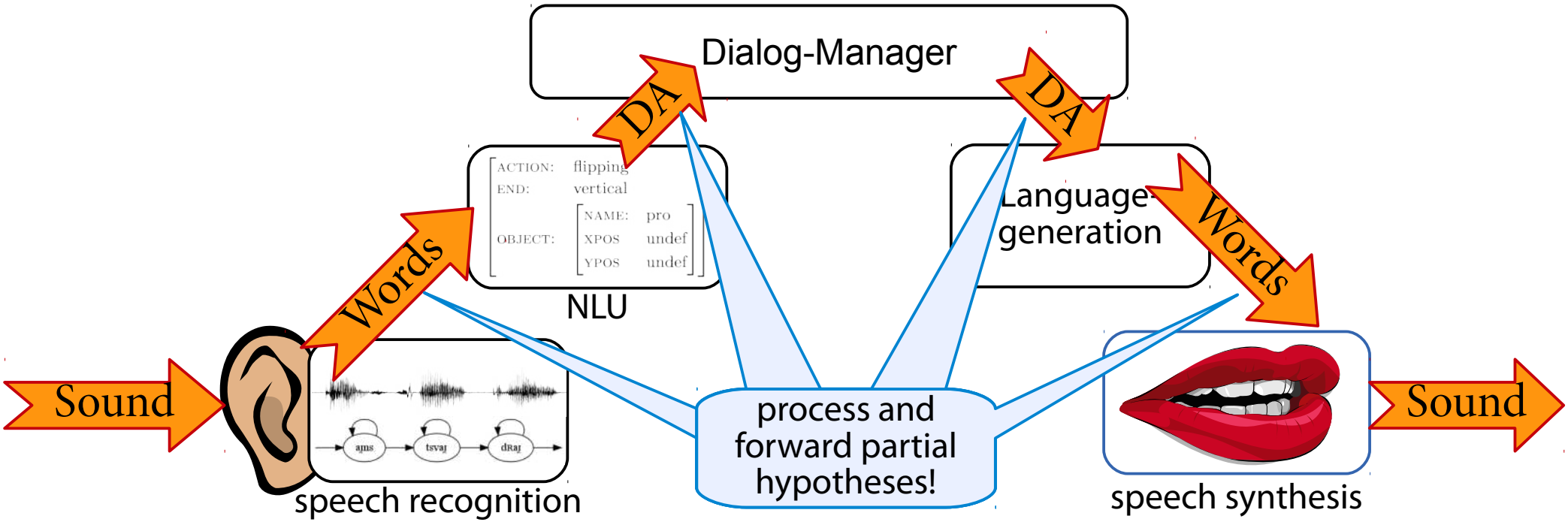
- Eingabe besteht aus minimalen Einheiten die stück-für-stück verarbeitet werden (Sprachaudio, Wörter, Gedanken, …)
- Ausgaben werden schon für Teileingaben erzeugt
- Eingabeeinheiten können zu größeren Ausgabeeinheiten zusammengefasst werden

Incremental Processing

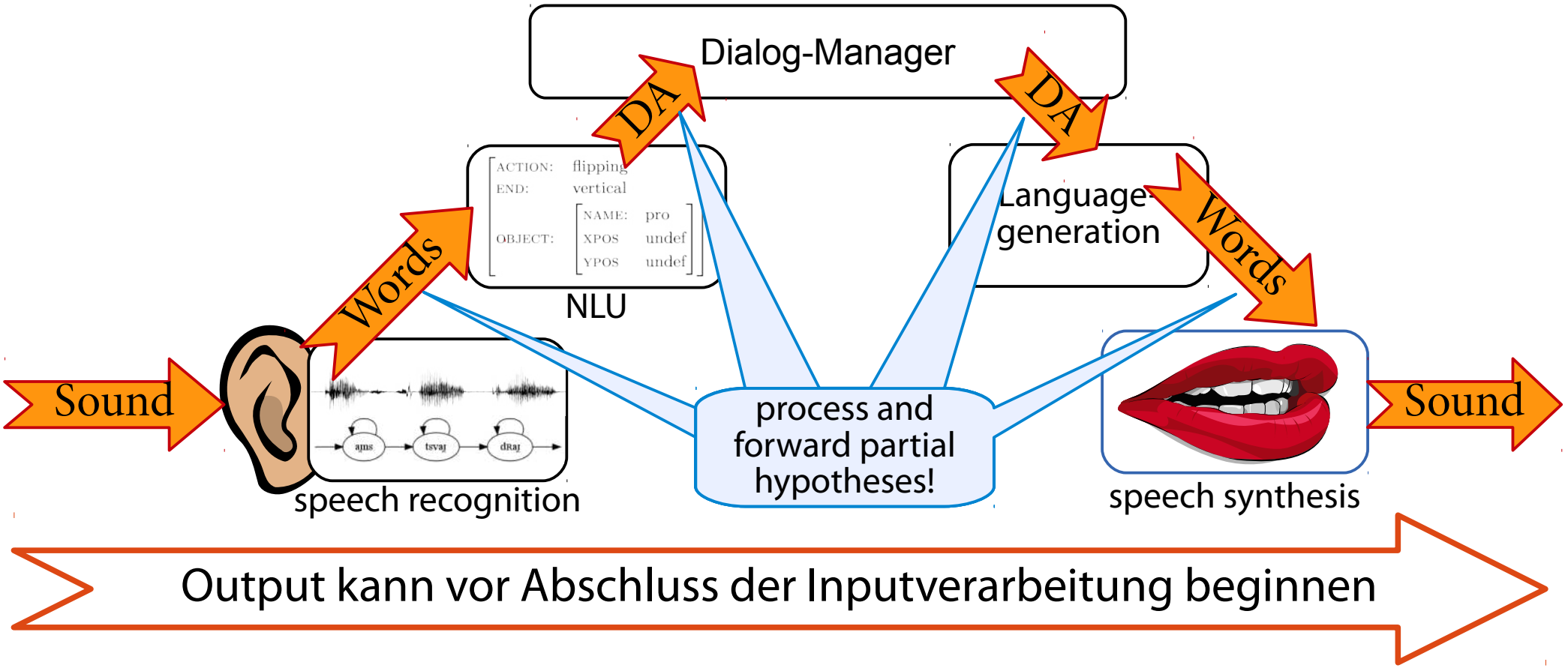


- Eingabe besteht aus minimalen Einheiten die stück-für-stück verarbeitet werden (Sprachaudio, Wörter, Gedanken, …)
- Ausgaben werden schon für Teileingaben erzeugt
- Eingabeeinheiten können zu größeren Ausgabeeinheiten zusammengefasst werden

ein modulares **inkrementelles** Dialogsystem



ein modulares **inkrementelles** Dialogsystem



Herausforderung Inkrementalität

four

[f O 6]

Herausforderung Inkrementalität

- Hypothesen basieren nur auf dem *was bereits bekannt ist*
  - mehr Kontext kann Ergebnisse beeinflussen
- example: speech recognition
  - input: [f O 6] → this sounds like “four”!
  - addition of [t i:] → together, this sounds like “fourty”!
  - what happens if [n] is next?

four

[f O 6]

Herausforderung Inkrementalität

- Hypothesen basieren nur auf dem *was bereits bekannt ist*
  - mehr Kontext kann Ergebnisse beeinflussen
- example: speech recognition
  - input: [f O 6] → this sounds like “four”!
  - addition of [t i:] → together, this sounds like “fourty”!
  - what happens if [n] is next?

four

[f O 6 t i:]

Herausforderung Inkrementalität

- Hypothesen basieren nur auf dem *was bereits bekannt ist*
  - mehr Kontext kann Ergebnisse beeinflussen
- example: speech recognition
  - input: [f O 6] → this sounds like “four”!
  - addition of [t i:] → together, this sounds like “fourty”!
  - what happens if [n] is next?

[f O 6 t i:]

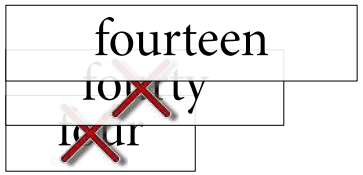
Herausforderung Inkrementalität

- Hypothesen basieren nur auf dem *was bereits bekannt ist*
  - mehr Kontext kann Ergebnisse beeinflussen
- example: speech recognition
  - input: [f O 6] → this sounds like “four”!
  - addition of [t i:] → together, this sounds like “fourty”!
  - what happens if [n] is next?

[f O 6 t i: n]

Herausforderung Inkrementalität

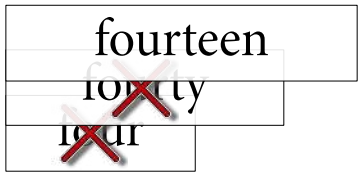
- Hypothesen basieren nur auf dem *was bereits bekannt ist*
  - mehr Kontext kann Ergebnisse beeinflussen
- example: speech recognition
  - input: [f O 6] → this sounds like “four”!
  - addition of [t i:] → together, this sounds like “fourty”!
  - what happens if [n] is next?



[f O 6 t i: n]

Herausforderung Inkrementalität

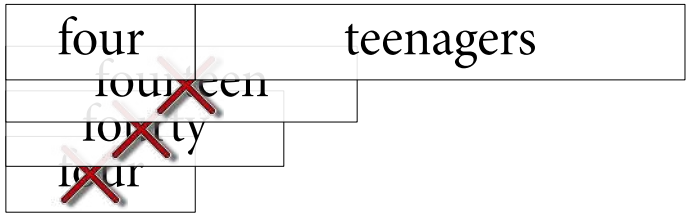
- Hypothesen basieren nur auf dem *was bereits bekannt ist*
  - mehr Kontext kann Ergebnisse beeinflussen
- example: speech recognition
  - input: [f O 6] → this sounds like “four”!
  - addition of [t i:] → together, this sounds like “fourty”!
  - what happens if [n] is next? then [EI dZ 6 z]?



[f O 6 t i: n EI dZ 6 z]

Herausforderung Inkrementalität

- Hypothesen basieren nur auf dem *was bereits bekannt ist*
  - mehr Kontext kann Ergebnisse beeinflussen
- example: speech recognition
  - input: [f O 6] → this sounds like “four”!
  - addition of [t i:] → together, this sounds like “fourty”!
  - what happens if [n] is next? then [EI dZ 6 z]?



[f O 6 t i: n EI dZ 6 z]

Herausforderung Inkrementalität

Baumann, Atterer & Schlangen (2009), McGraw & Gruenstein (2012)

Herausforderung Inkrementalität

- Hypothesen basieren nur auf dem *was bereits bekannt ist*
  - mehr Kontext kann Ergebnisse beeinflussen
- Notwendigkeit, später Änderungen zu ermöglichen
- lohnt es sich trotzdem noch?

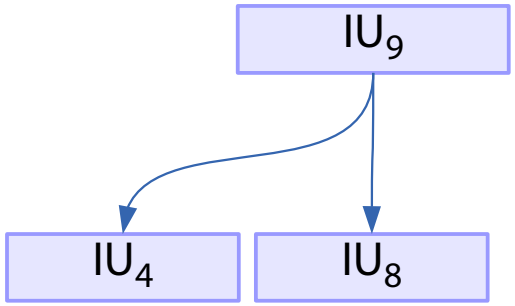
Baumann, Atterer & Schlangen (2009), McGraw & Gruenstein (2012)

Architektur-Grundlage

Incremental Unit model (Schlangen & Skantze 2009, Baumann 2013)

- atomar = minimale Menge an Information auf der gegebenen Abstraktionsebene

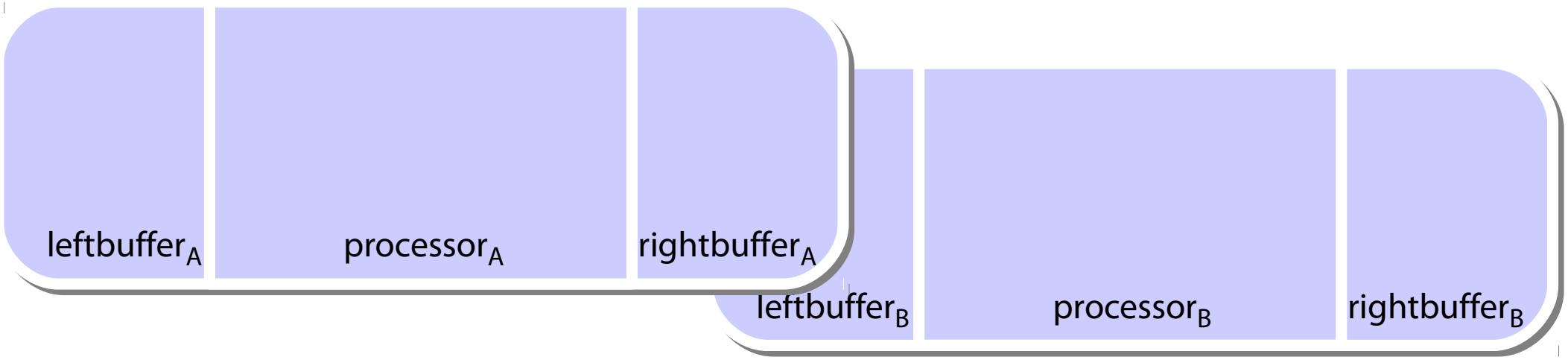
- Verknüpfung mit zugehörigen Einheiten: Gesamtnetz aller Einheiten stellt den Informationszustand des Systems dar



- linking erfasst **Dependenzen** zwischen Einheiten
  - allows to drill into genealogy of the input (white-box approach)
    - semantics of "umfahren"? → inspect prosodic realization!
  - allows to track changes to **revise** one's previous output hypothesis

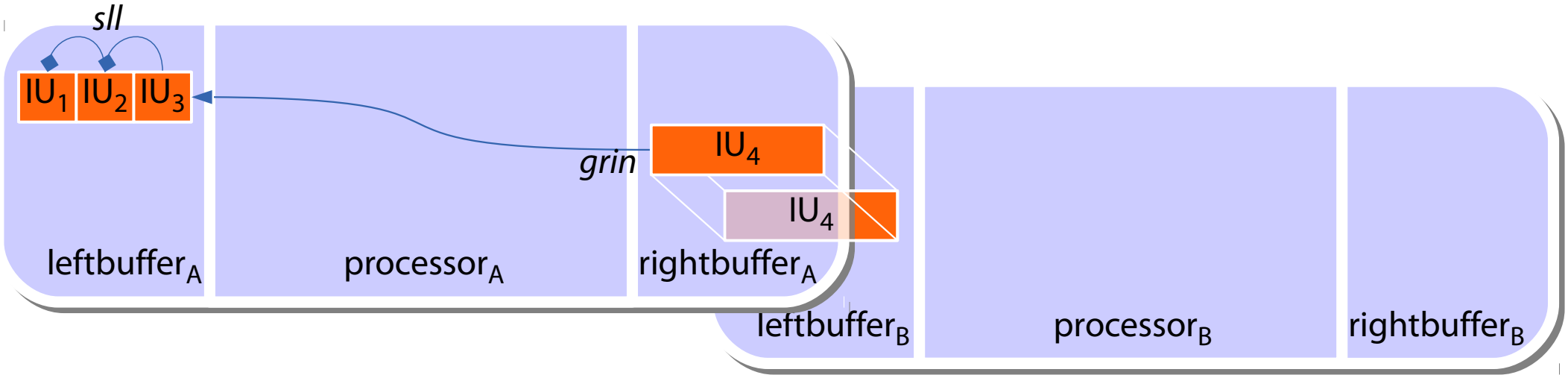
Processing modules

- processing modules are connected via buffers



Processing modules

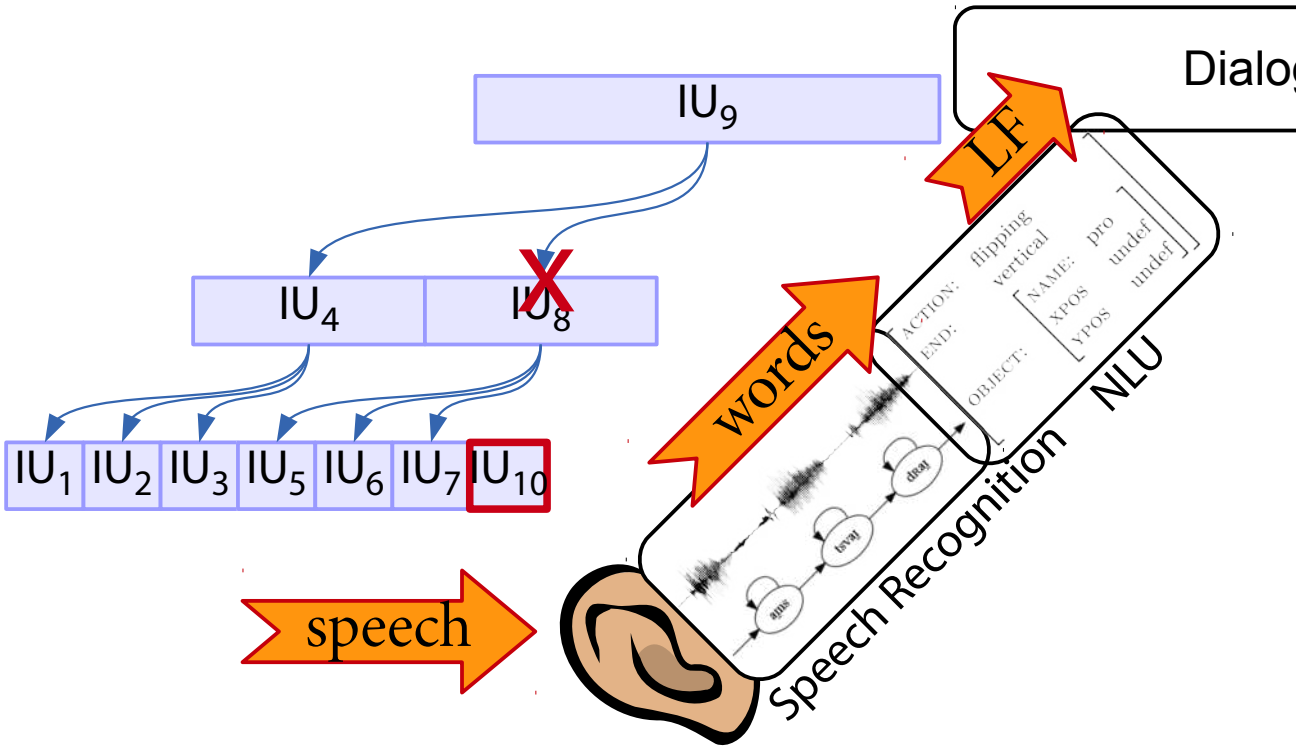
- processing modules are connected via buffers
- buffers contain incremental units (IUs)



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

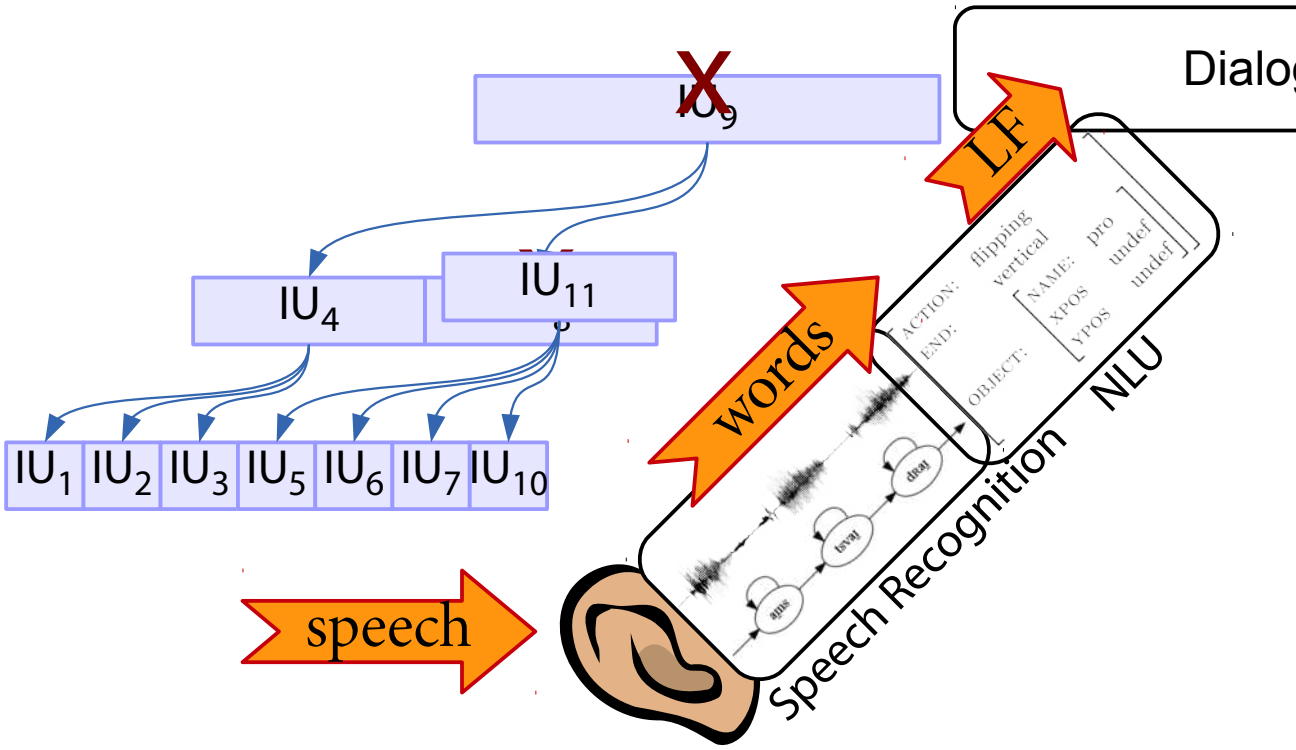
Incremental Unit Network

- belief changes reflected by changes in the network
  - more audio input arrives
  - word hypothesis is revoked …



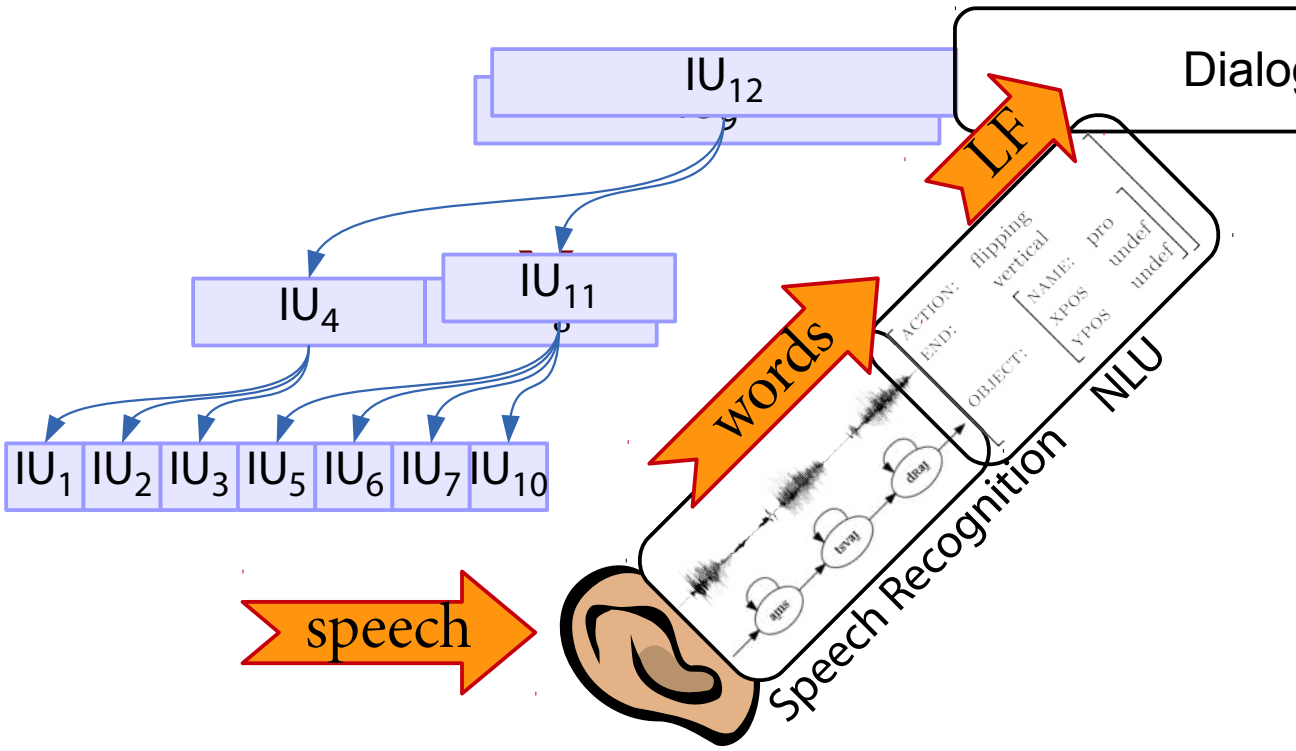
Incremental Unit Network

- belief changes reflected by changes in the network
  - more audio input arrives
  - word hypothesis is revoked and replaced by a different one

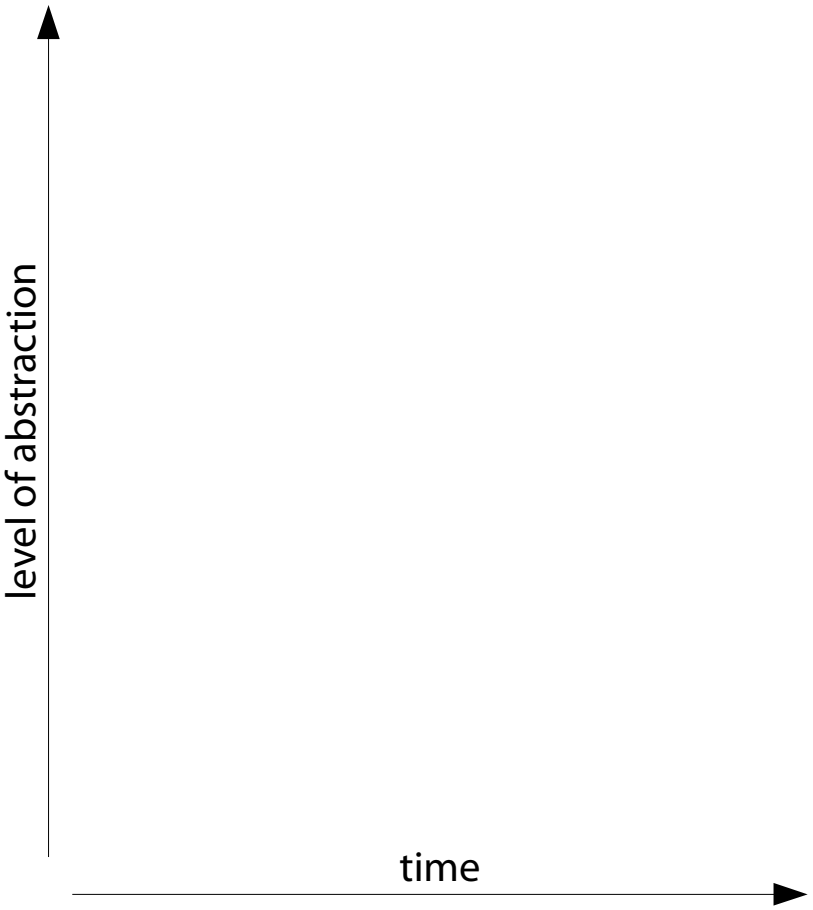


Incremental Unit Network

- belief changes reflected by changes in the network
  - more audio input arrives
  - word hypothesis is revoked and replaced by a different one
  - changes trickle up in the system



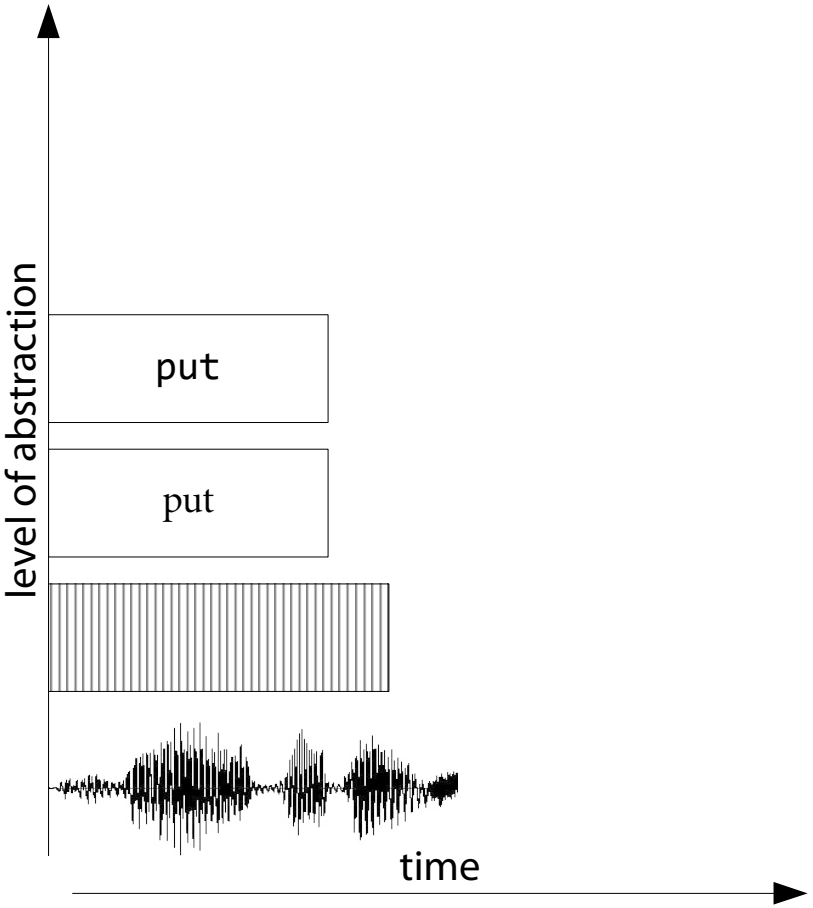
Consuming input incrementally



input side

Baumann (2013)

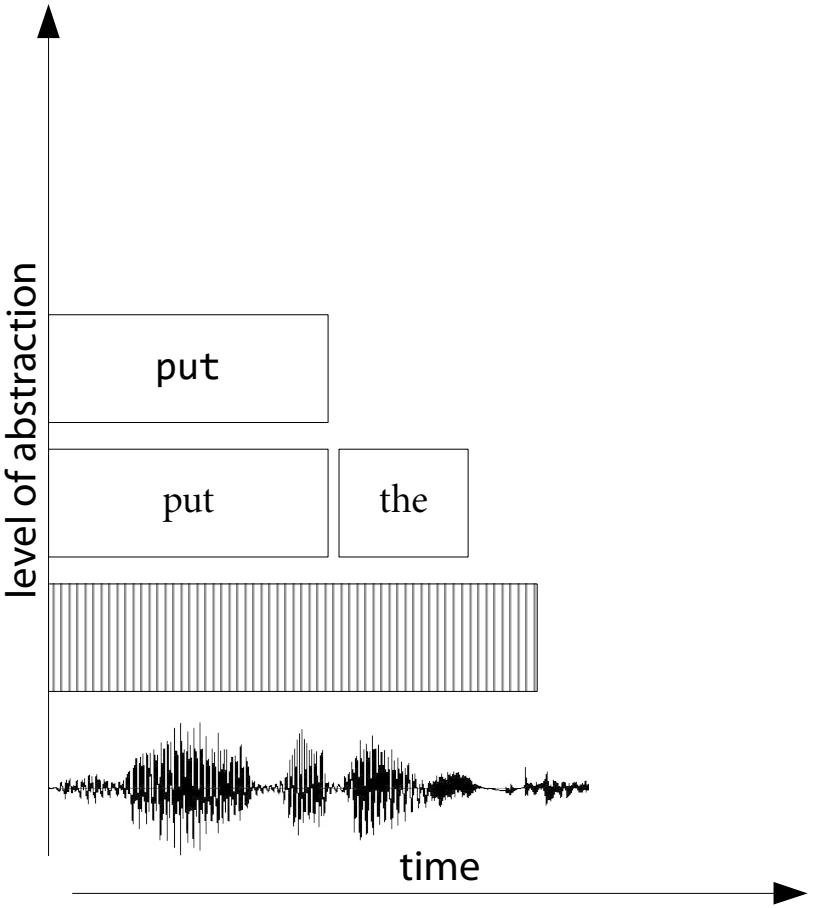
Consuming input incrementally



input side

Baumann (2013)

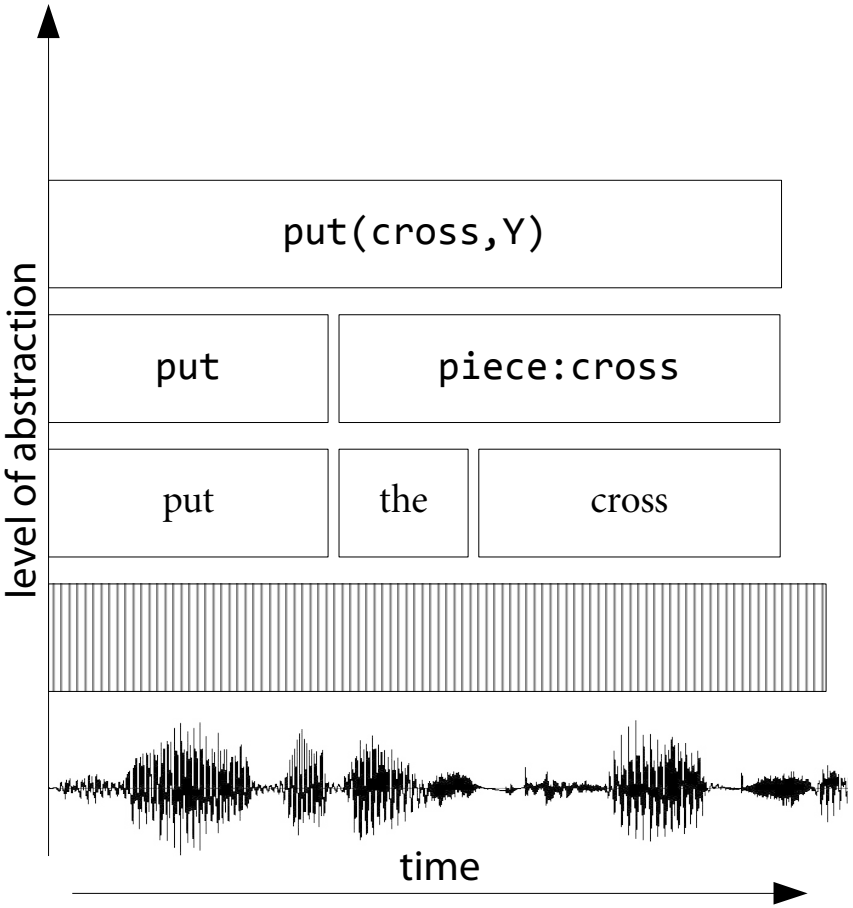
Consuming input incrementally



input side

Baumann (2013)

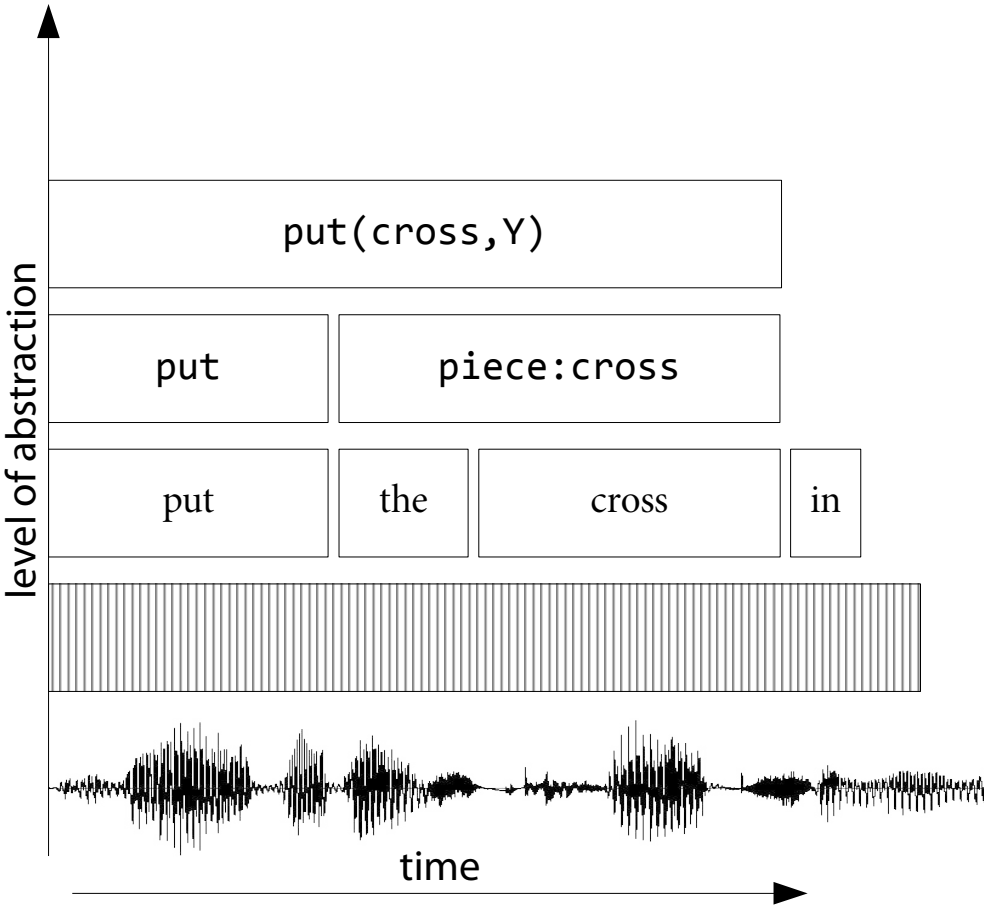
Consuming input incrementally



input side

Baumann (2013)

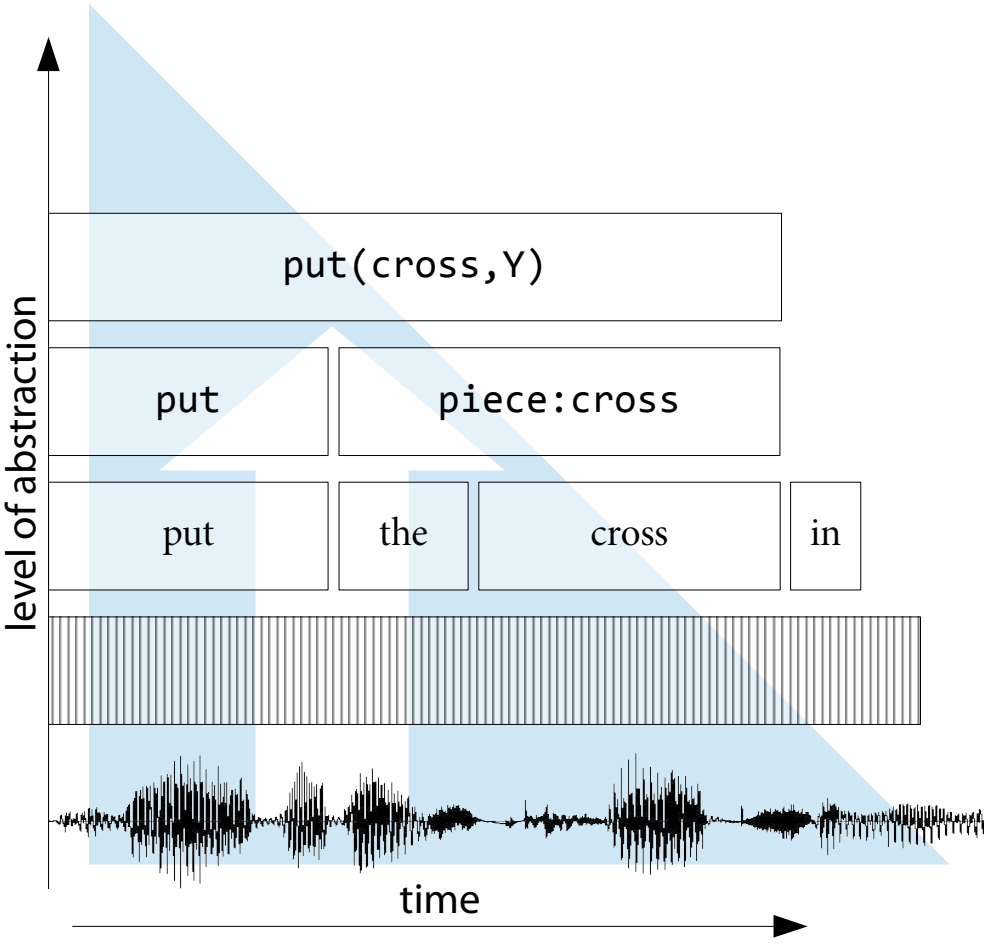
Consuming input incrementally



input side

Baumann (2013)

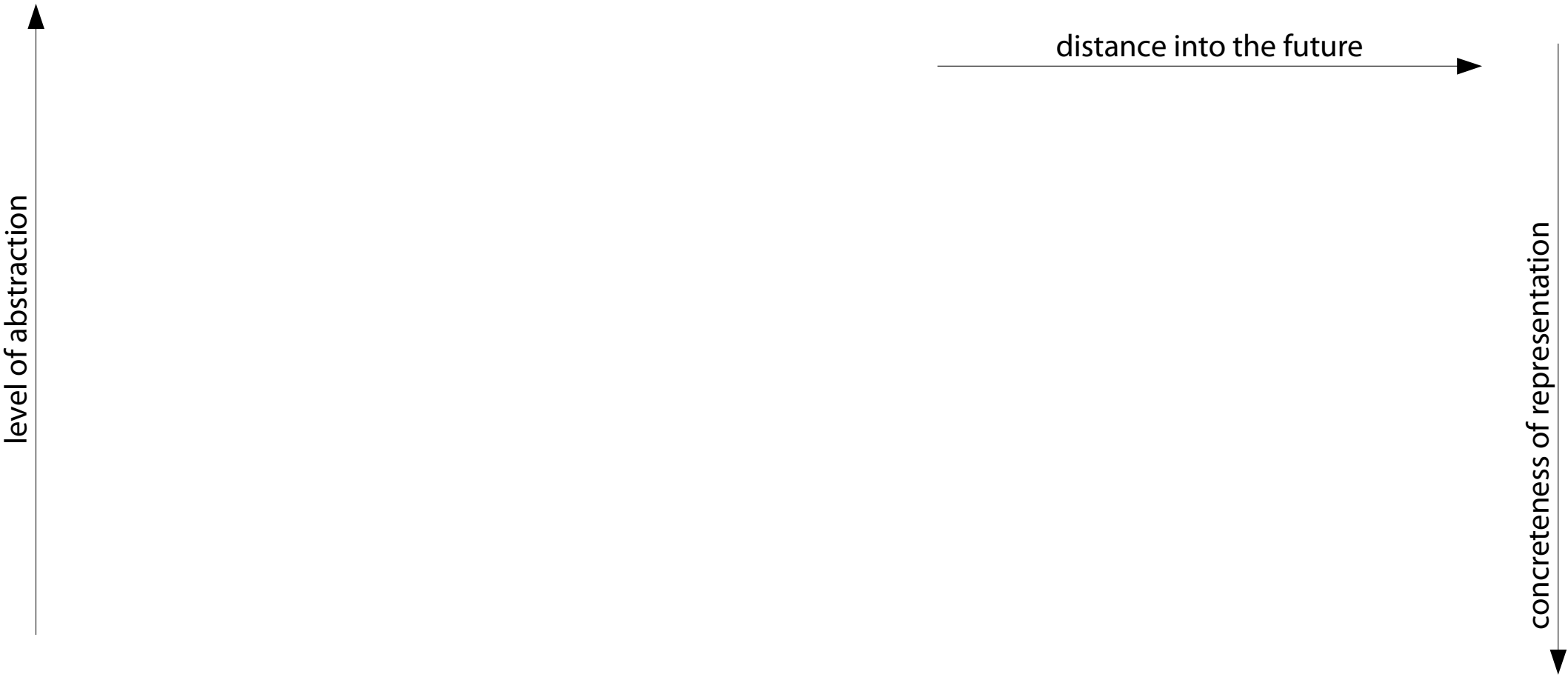
Consuming input incrementally



input side

Baumann (2013)

Producing output just-in-time



output side

Baumann (2013)

Producing output just-in-time

output side

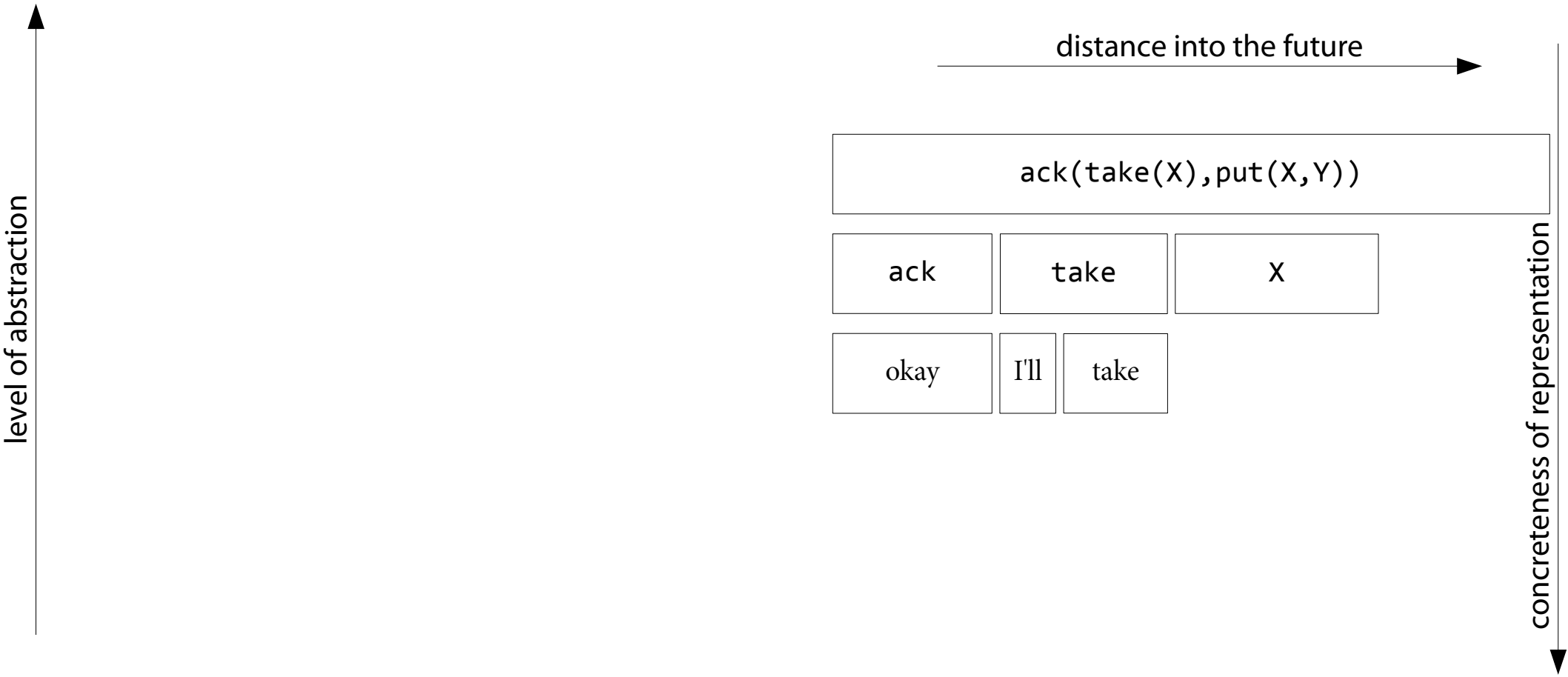
Baumann (2013)

Producing output just-in-time

output side

Baumann (2013)

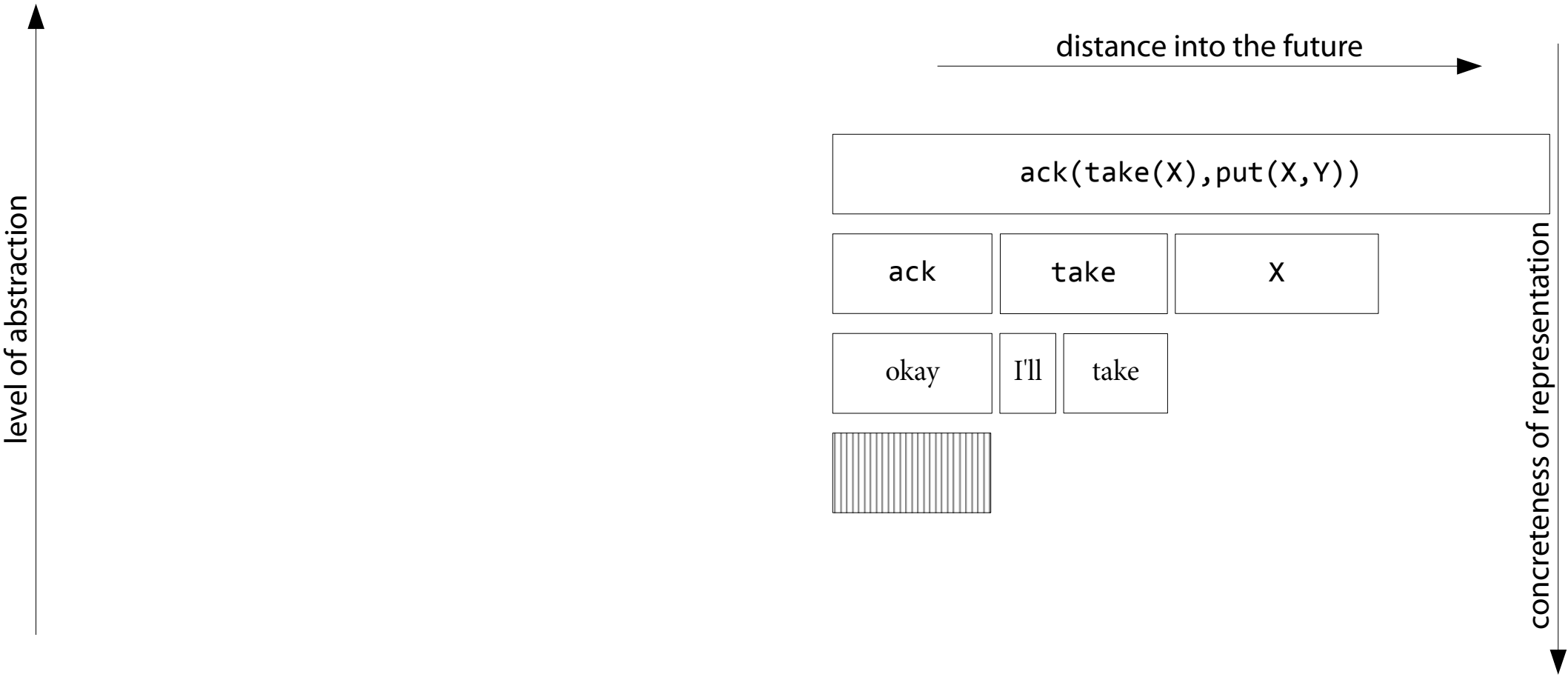
Producing output just-in-time



output side

Baumann (2013)

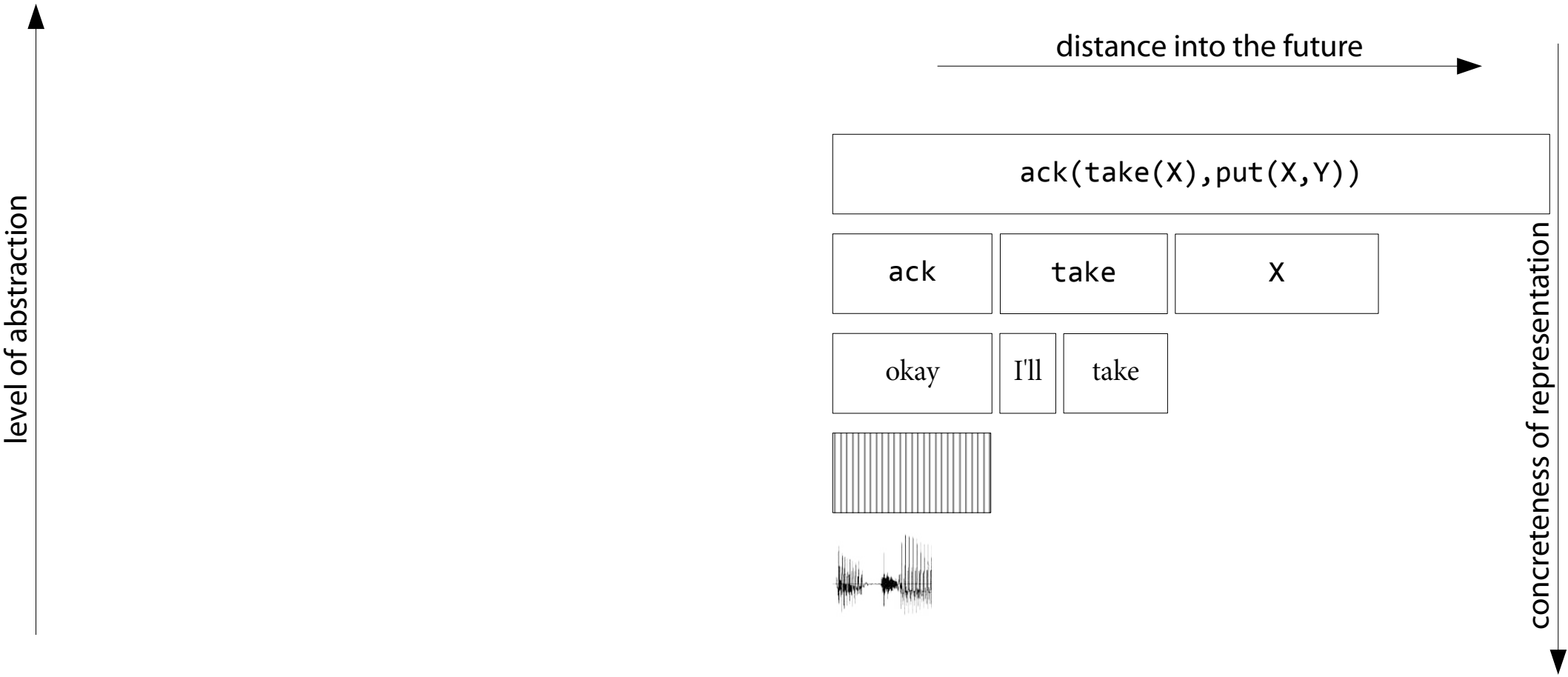
Producing output just-in-time



output side

Baumann (2013)

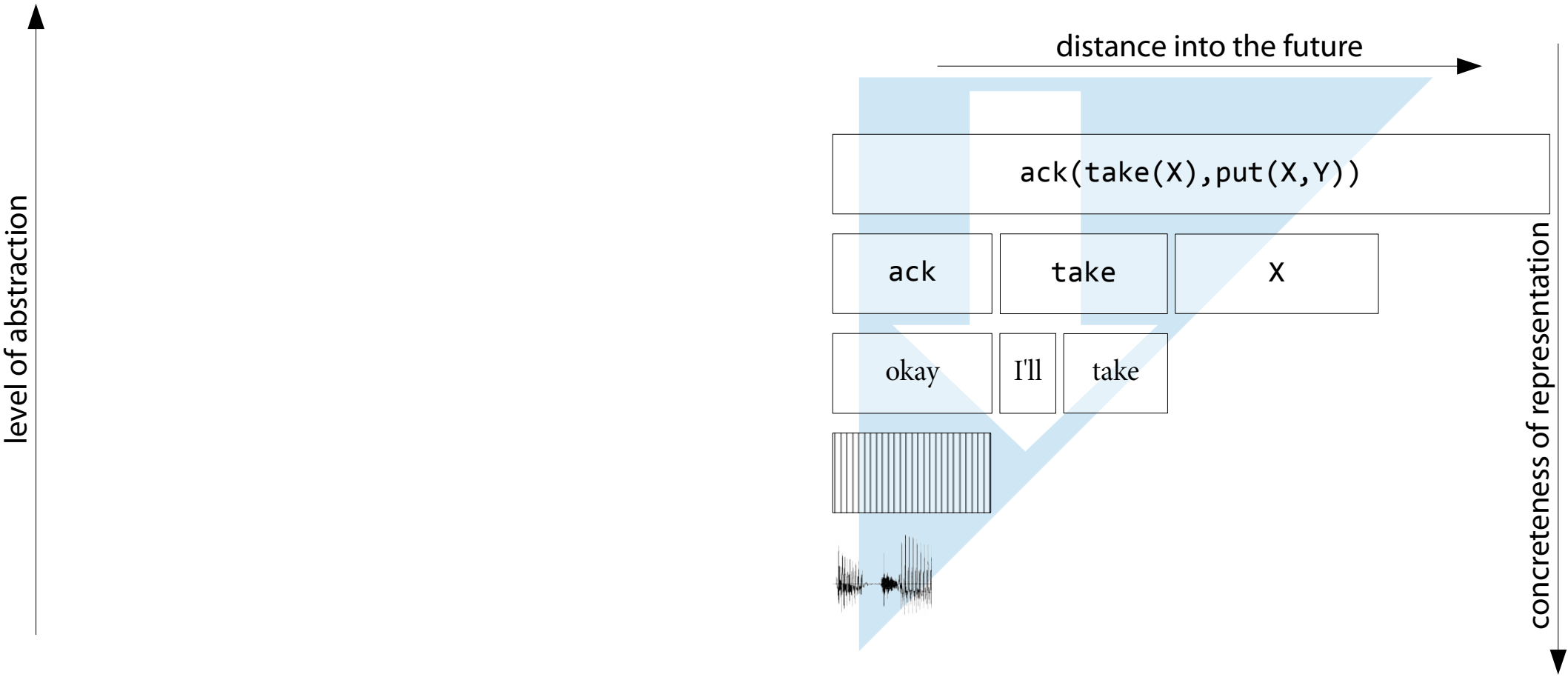
Producing output just-in-time



output side

Baumann (2013)

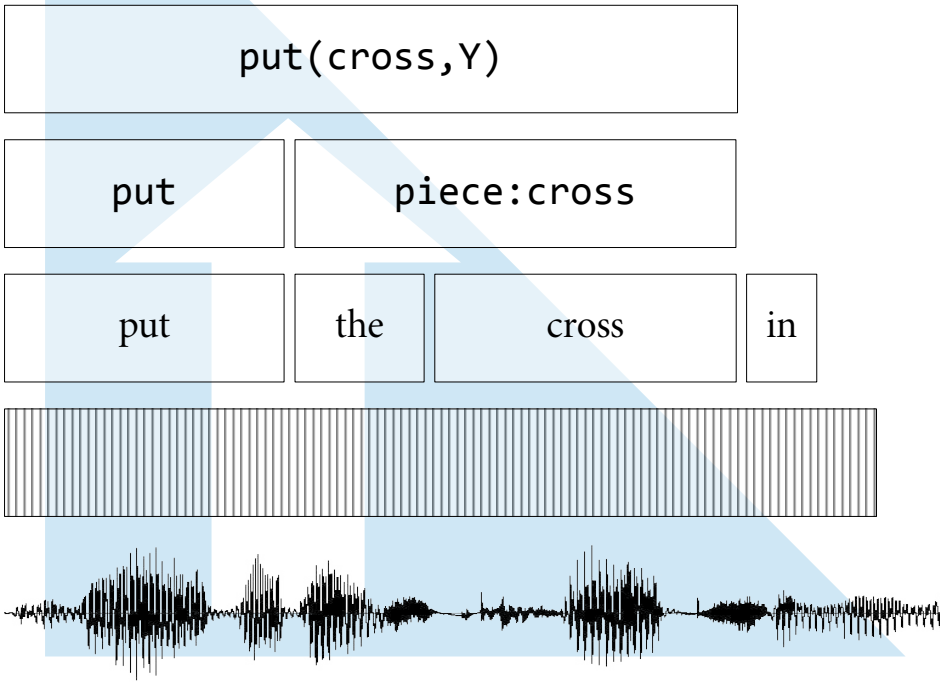
Producing output just-in-time

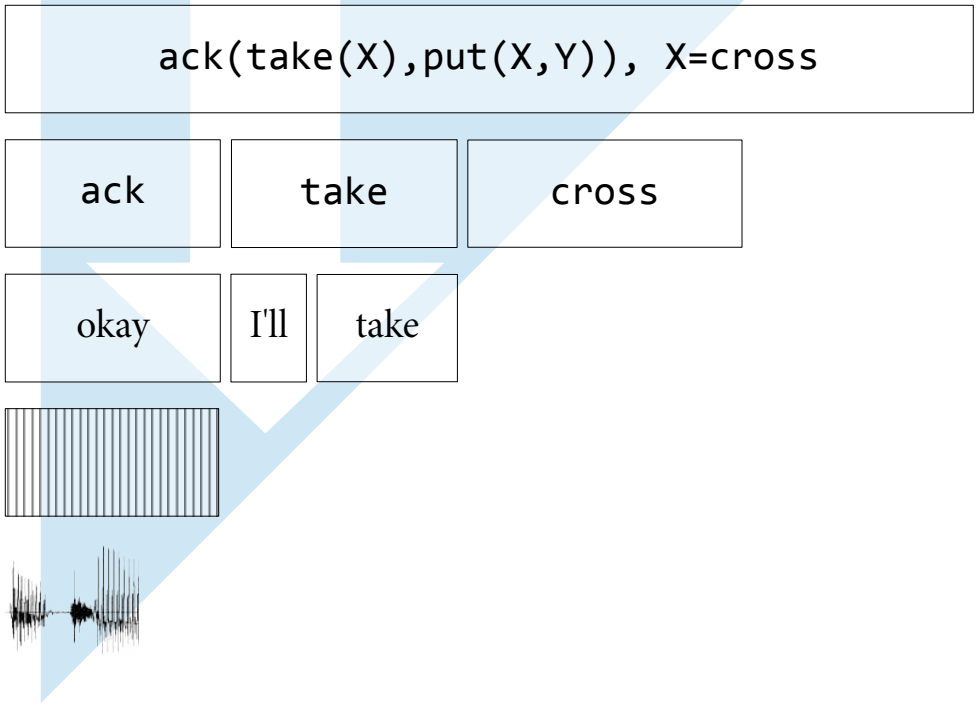


output side

Baumann (2013)

A data model for incremental just-in-time processing



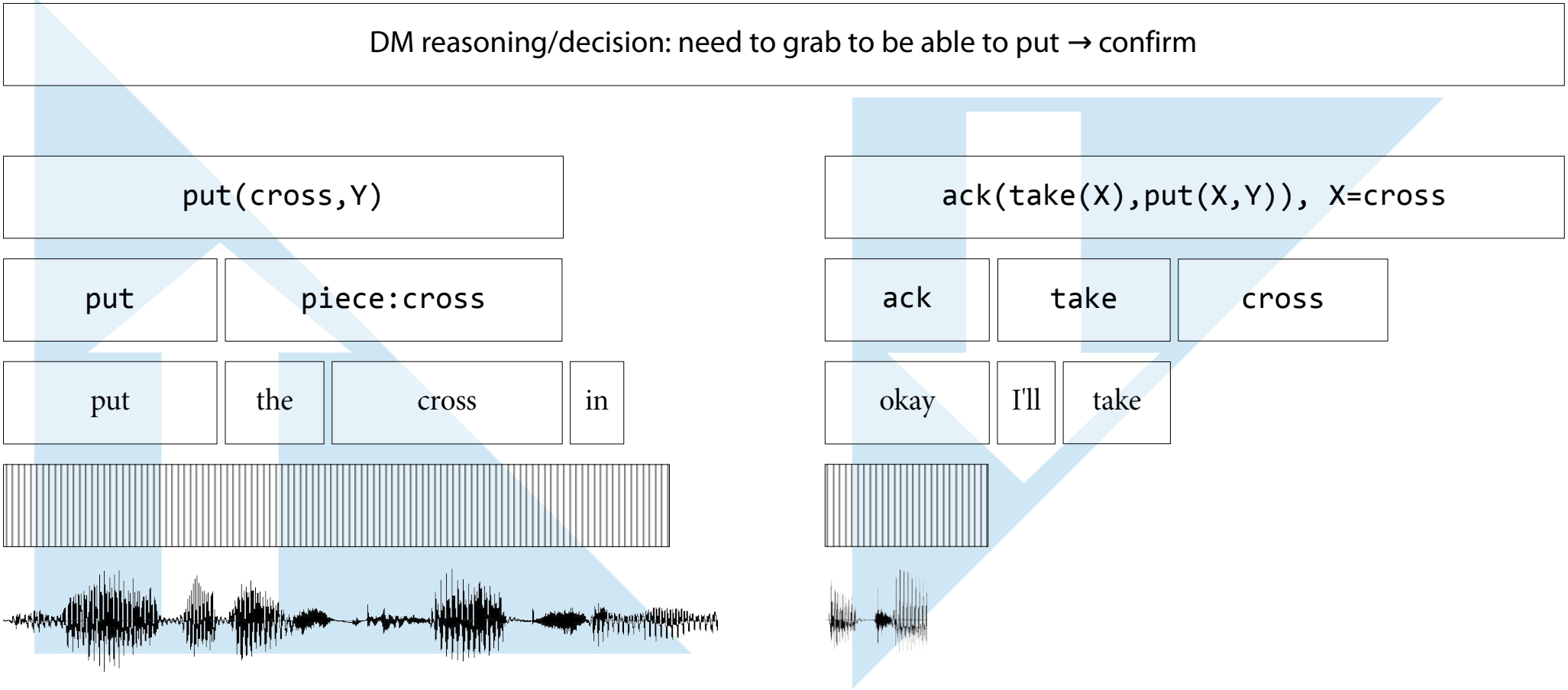


input side

output side

Baumann (2013)

A data model for incremental just-in-time processing



input side

output side

Baumann (2013)

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)

~~Drei~~ ein paar Beispiele für inkrementelle Verarbeitung

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

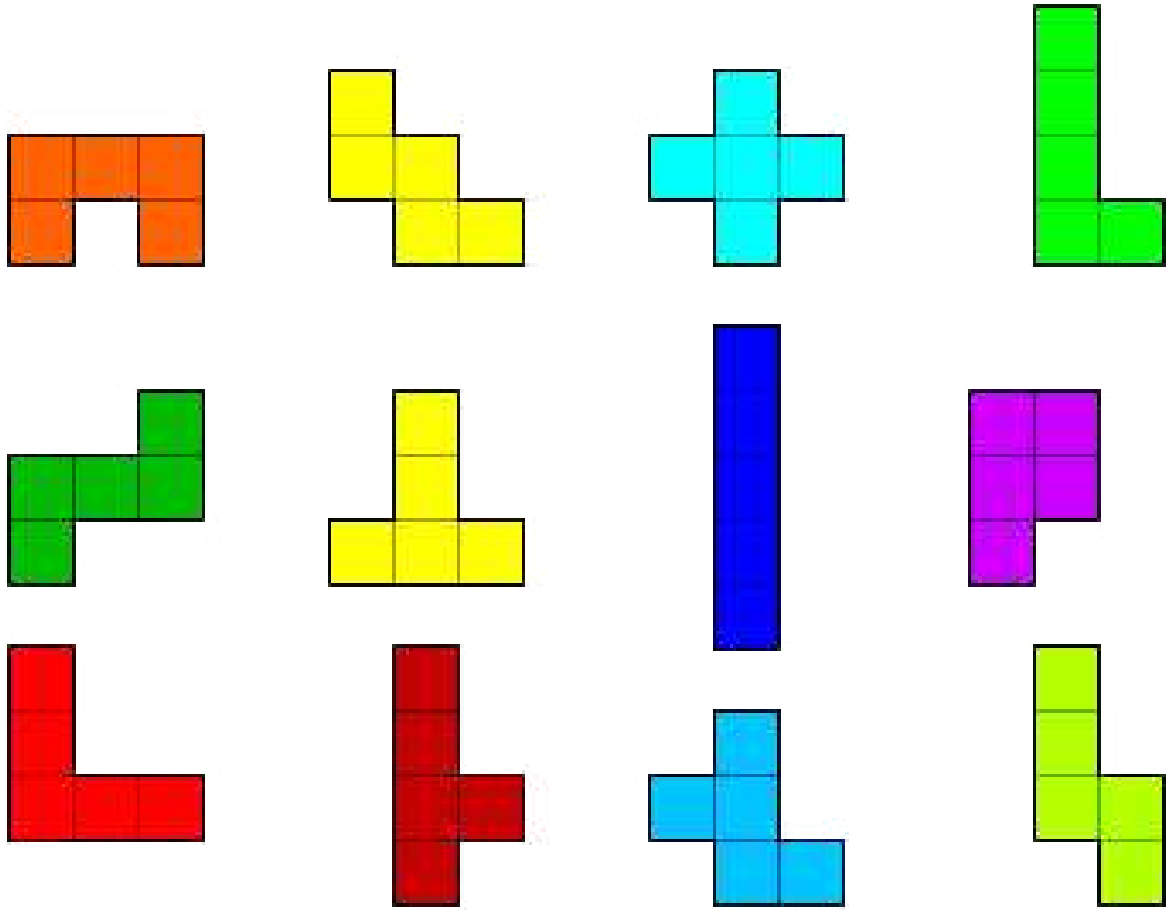
for the „micro-domain principle“ see (Edlund et al, Speech Communication 2008).

Feedback and sub-turn interaction

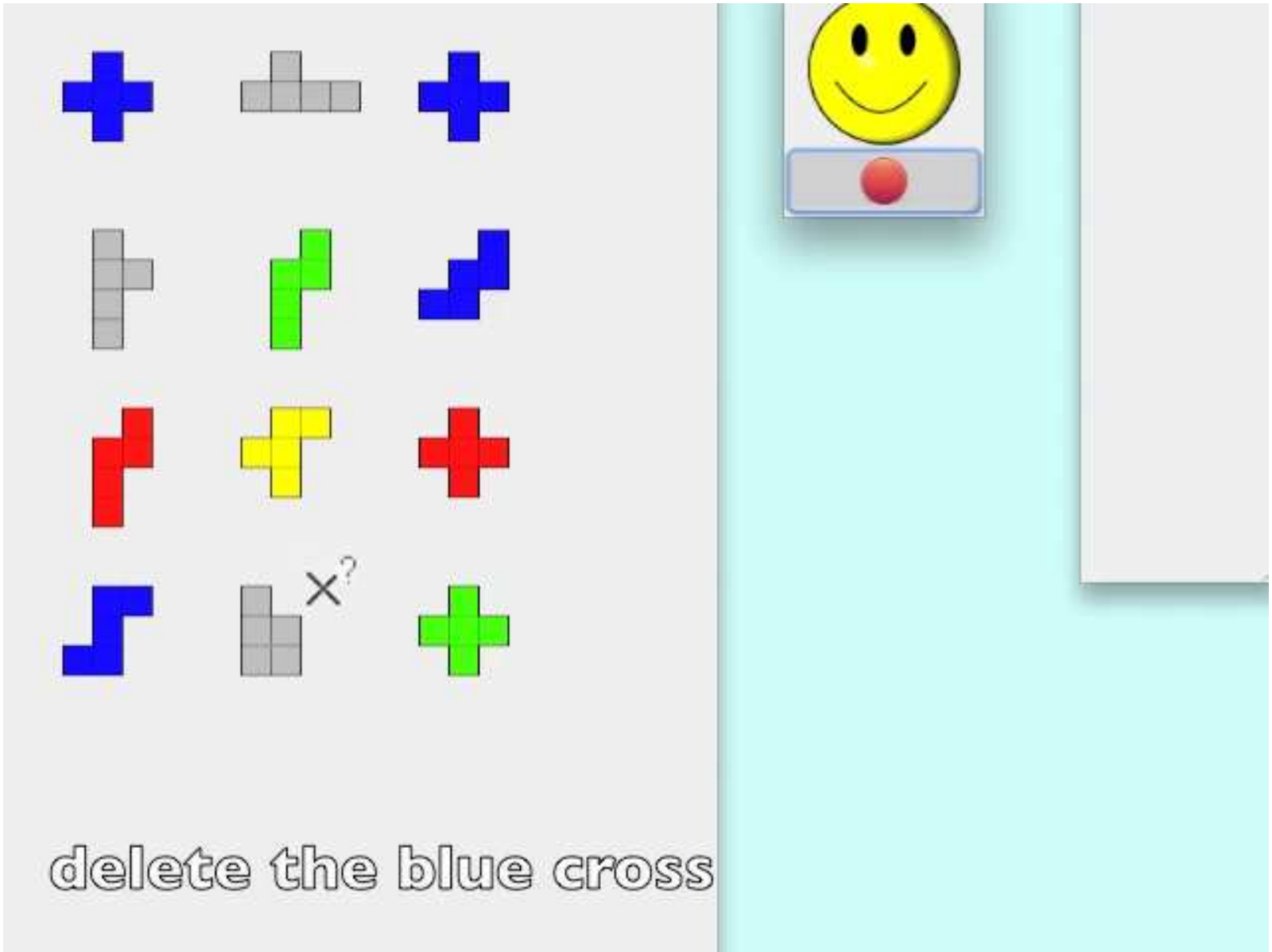
- Humans use feedback to signal state of understanding
  - often within a very tight *feedback loop*
  - incremental processing allows to tighten this feedback loop
  - in the video (to follow): visual feedback during the utterance
- Human reaction time (and type of reaction) depends on pragmatic completeness 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

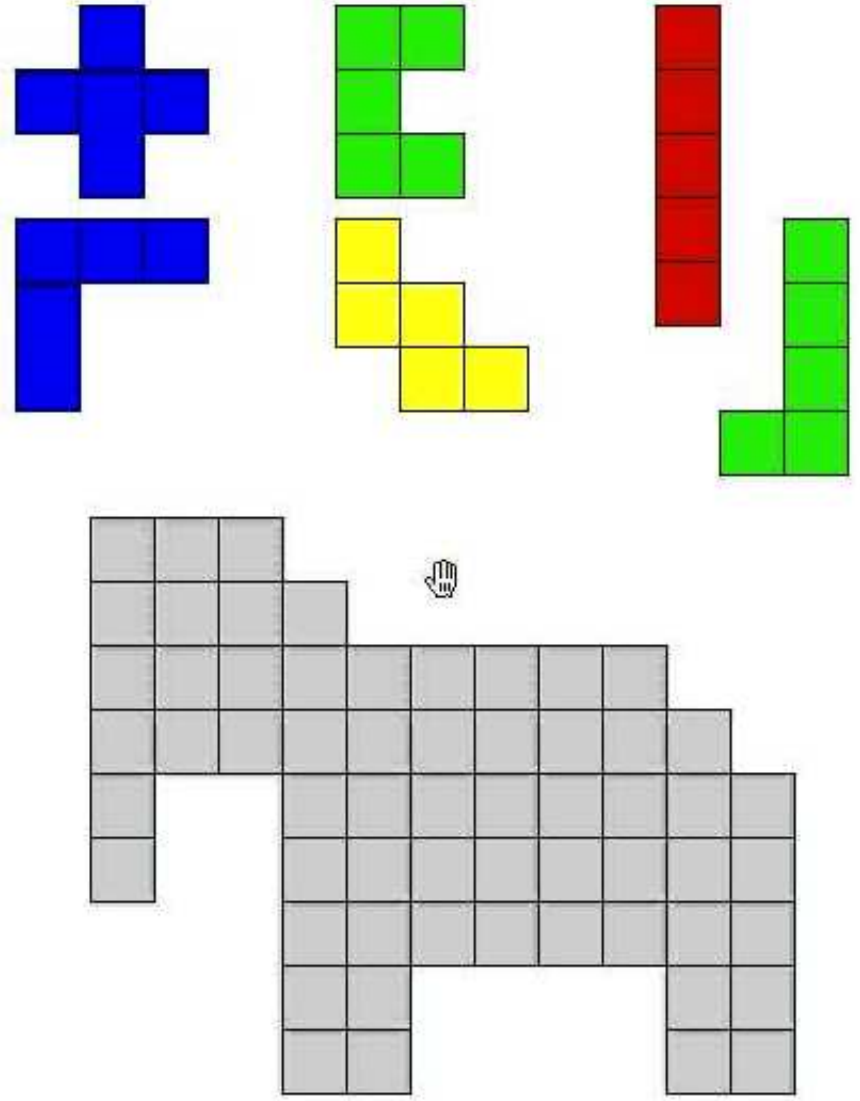


Feedback and sub-turn interaction

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

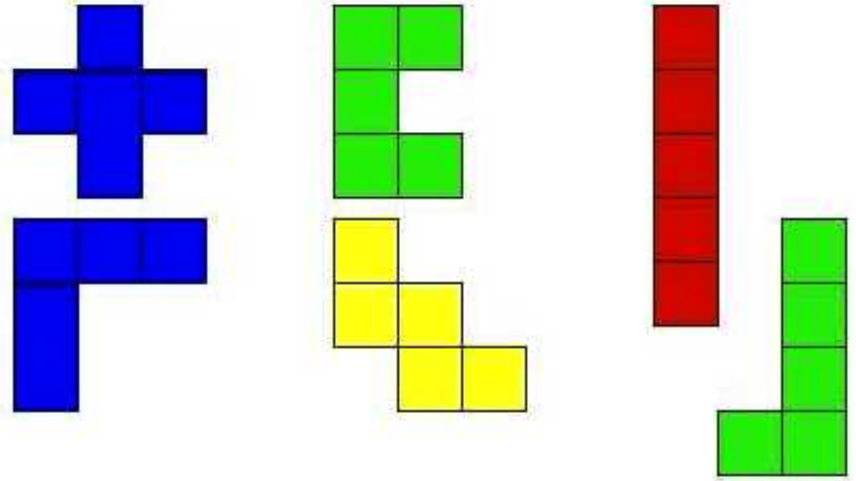
- overhearer study showed significantly better rated interactions over a baseline system
  - despite the differences between the systems being very subtle
  - small difference in behaviour → large difference in impression

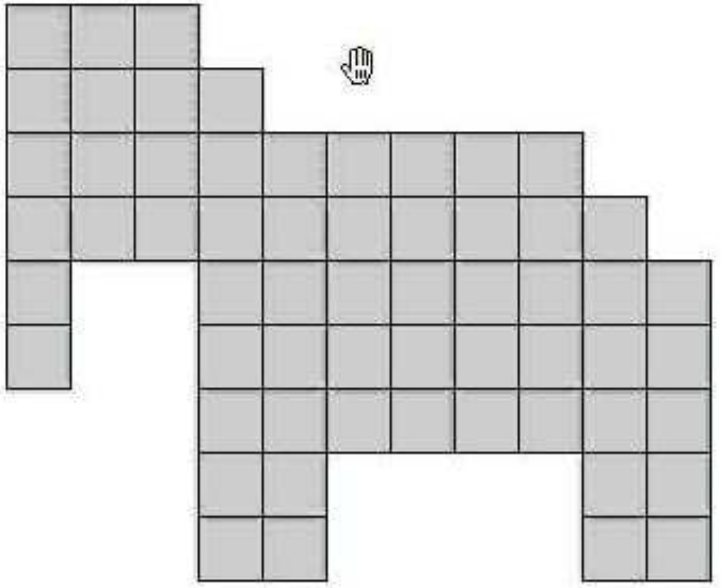
Playing puzzle games



Playing puzzle games

- how to puzzle an elefant?
- game alternates between
  - selecting puzzle piece
  - and placing it on the board



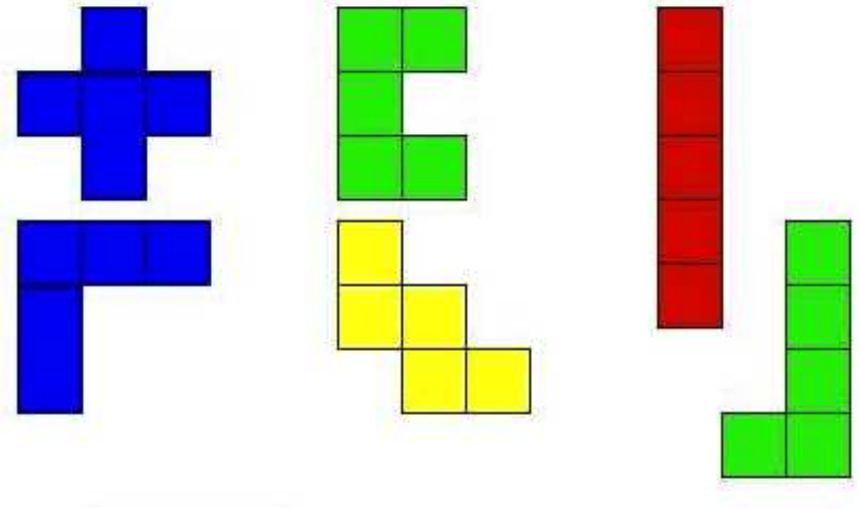


→ main challenge:
referring expressions

→ move complexity into
the interaction loop:
steering instead of naming

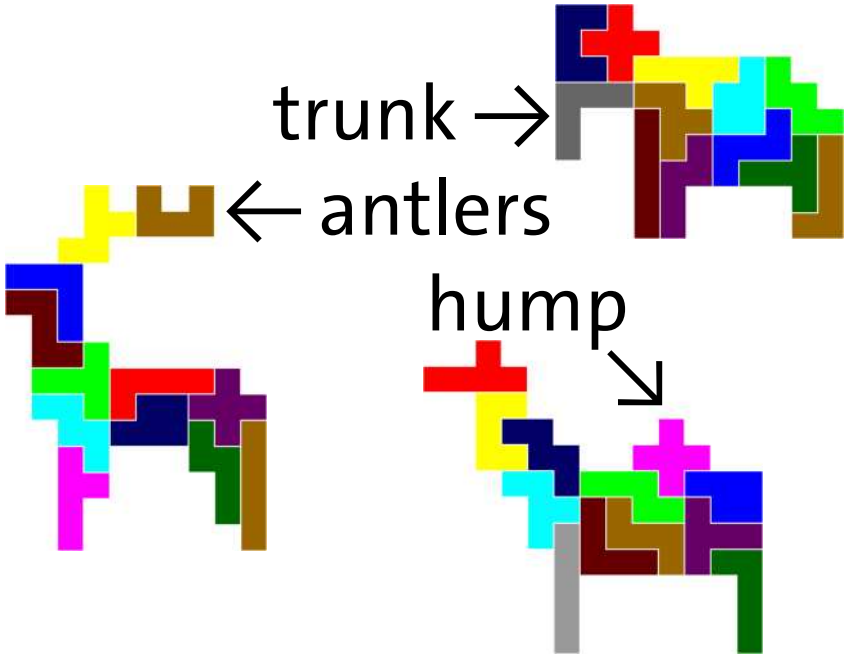
Playing puzzle games

- how to puzzle an elefant?
- game alternates between
  - selecting puzzle piece
  - and placing it on the board



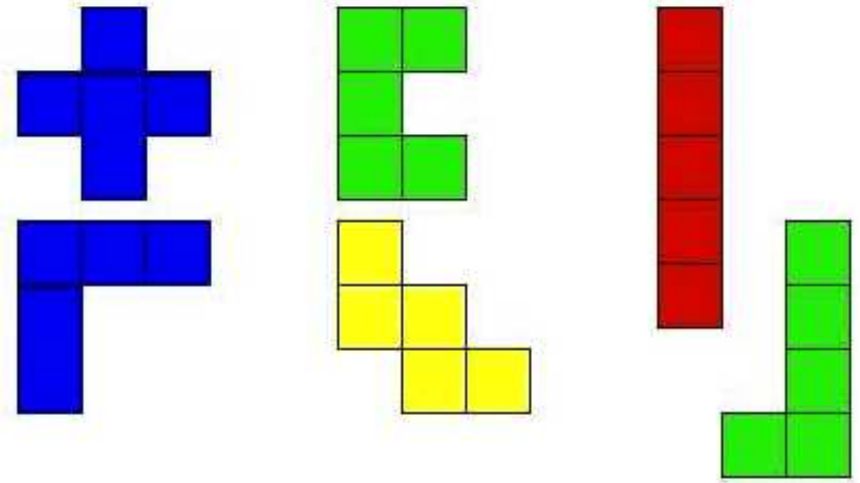
→ main challenge:
referring expressions

→ move complexity into
the interaction loop:
steering instead of naming



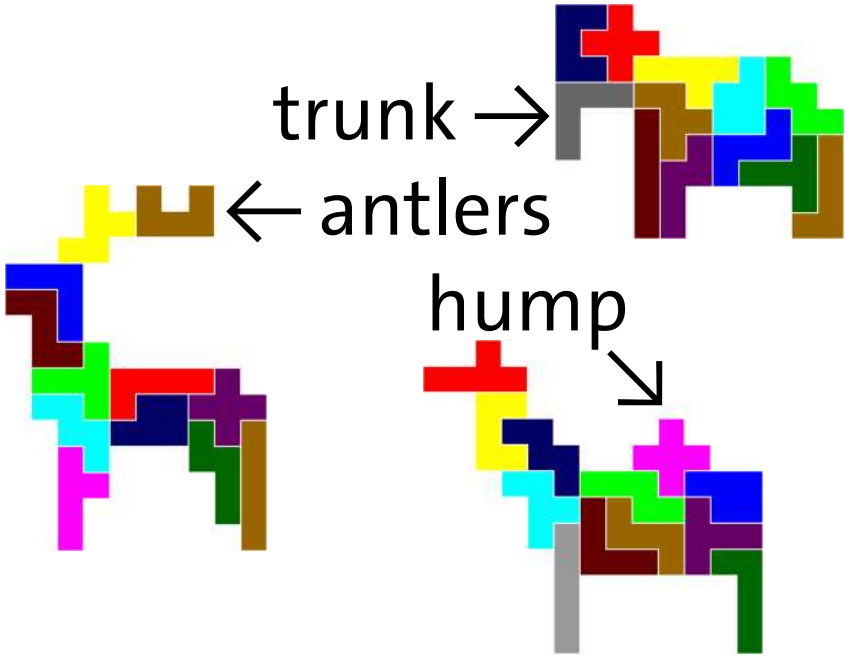
Playing puzzle games

- how to puzzle an elefant?
- game alternates between
  - selecting puzzle piece
  - and placing it on the board



→ main challenge:
referring expressions

→ move complexity into
the interaction loop:
steering instead of naming



Baumann et al. (2013)

Affordance as a driving principle

Jenkins, Journal of Scientific Psychology, pp. 34-45, Dec. 2008.

Affordance as a driving principle

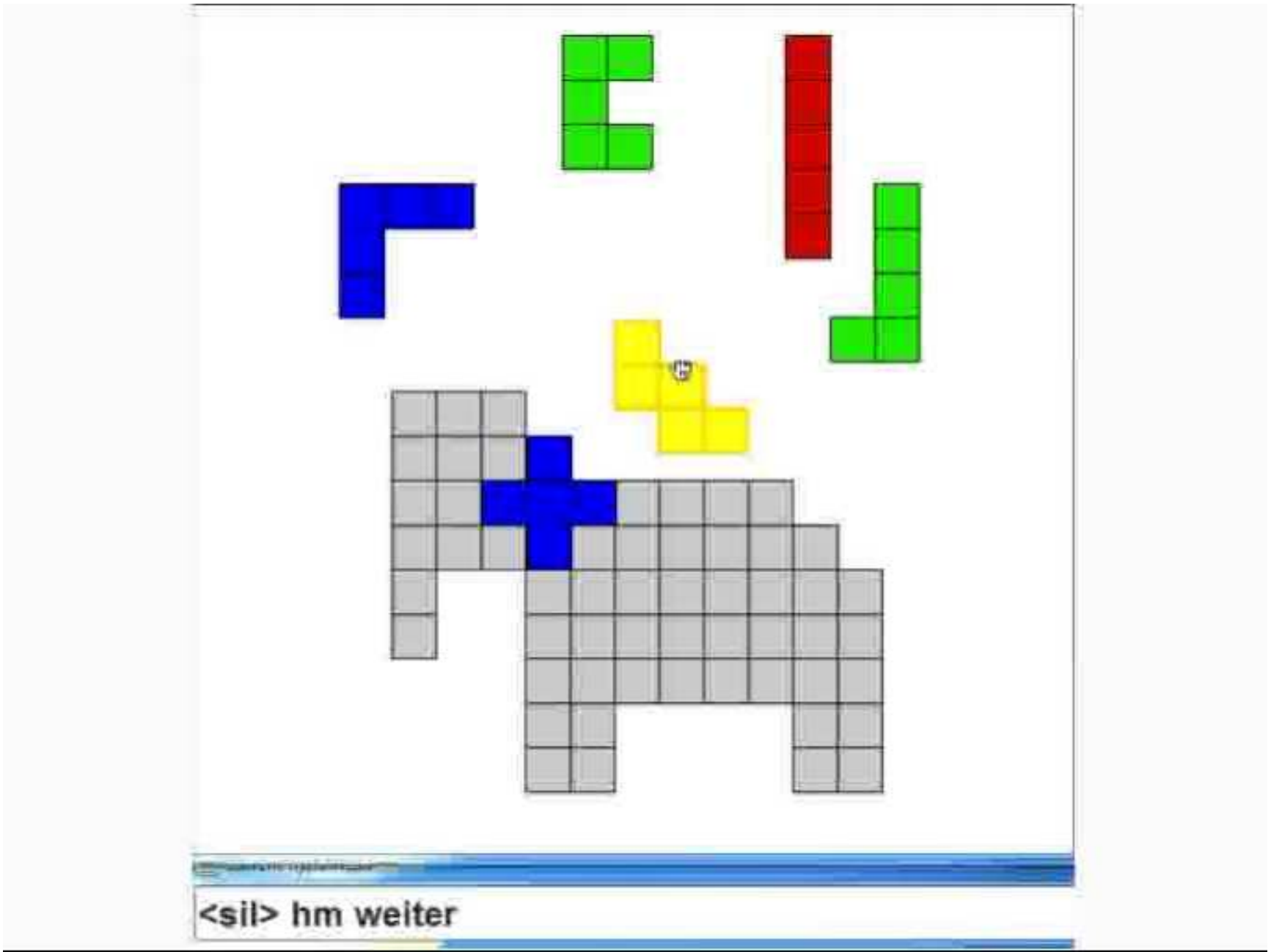
- affordances: conventionalized attribute-meaning pairs that manifest possibilities of interaction
  - doors afford to be opened
  - questions afford to be answered („where should I put the piece?“)
- *motion* manifests the possibility of interacting with the motion itself (*steering*)
  - steering (in 2D) is comparatively easy
  - to keep up the steering metaphor, the system must react to commands without noticeable delay

Jenkins, Journal of Scientific Psychology, pp. 34-45, Dec. 2008.

Example Video

please focus on:
- swiftness of system reactions
- error recovery
(ASR results shown at the bottom)

Steering metaphor in interaction



Experimental Evaluation

Baumann et al. (2013)

Experimental Evaluation

system was tested with/without immediate motion after the positioning question

- all users react to the affordance of motion (i.e., give steering commands)
- significantly faster task completion
- user questionnaire indicates advantage for affordance of motion (rated more transparent and reactive)

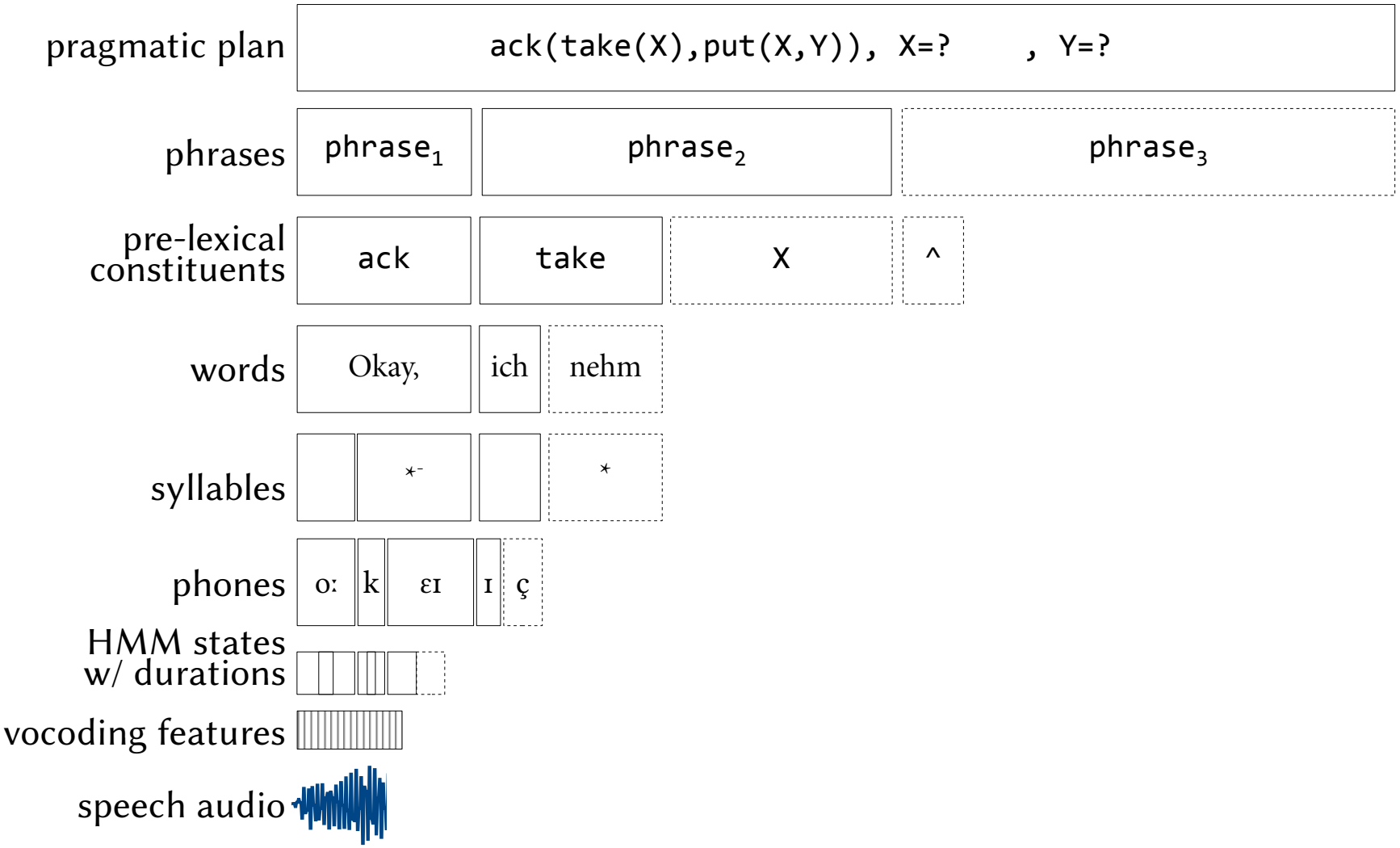
→ **ASR errors inhibit understanding, but not interaction!**

Baumann et al. (2013)

Affordances: Take-away Message

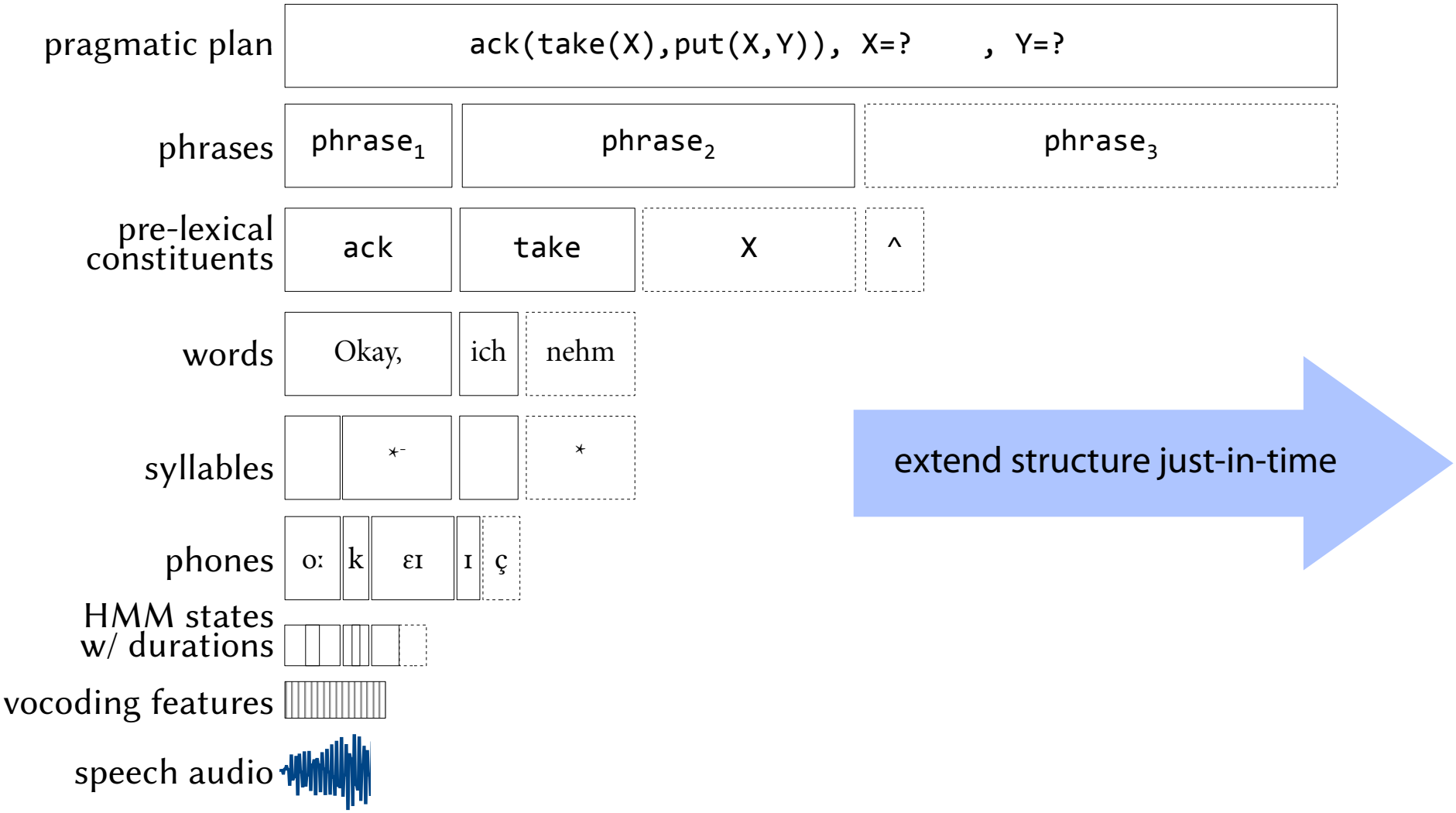
- move complexity where it hurts least / is manageable most easily
  - ask/act often in small steps → incrementally!
- think about what you propose to a user / what affordances are opened up
  - the relative strength of concurrent affordances
    - should the system act, ask, or do both?
    - how about the ordering of these?
  - the *ease of use* of affordances (e.g. steering is easy)

Incremental Speech Generation and Synthesis (HMM-based)



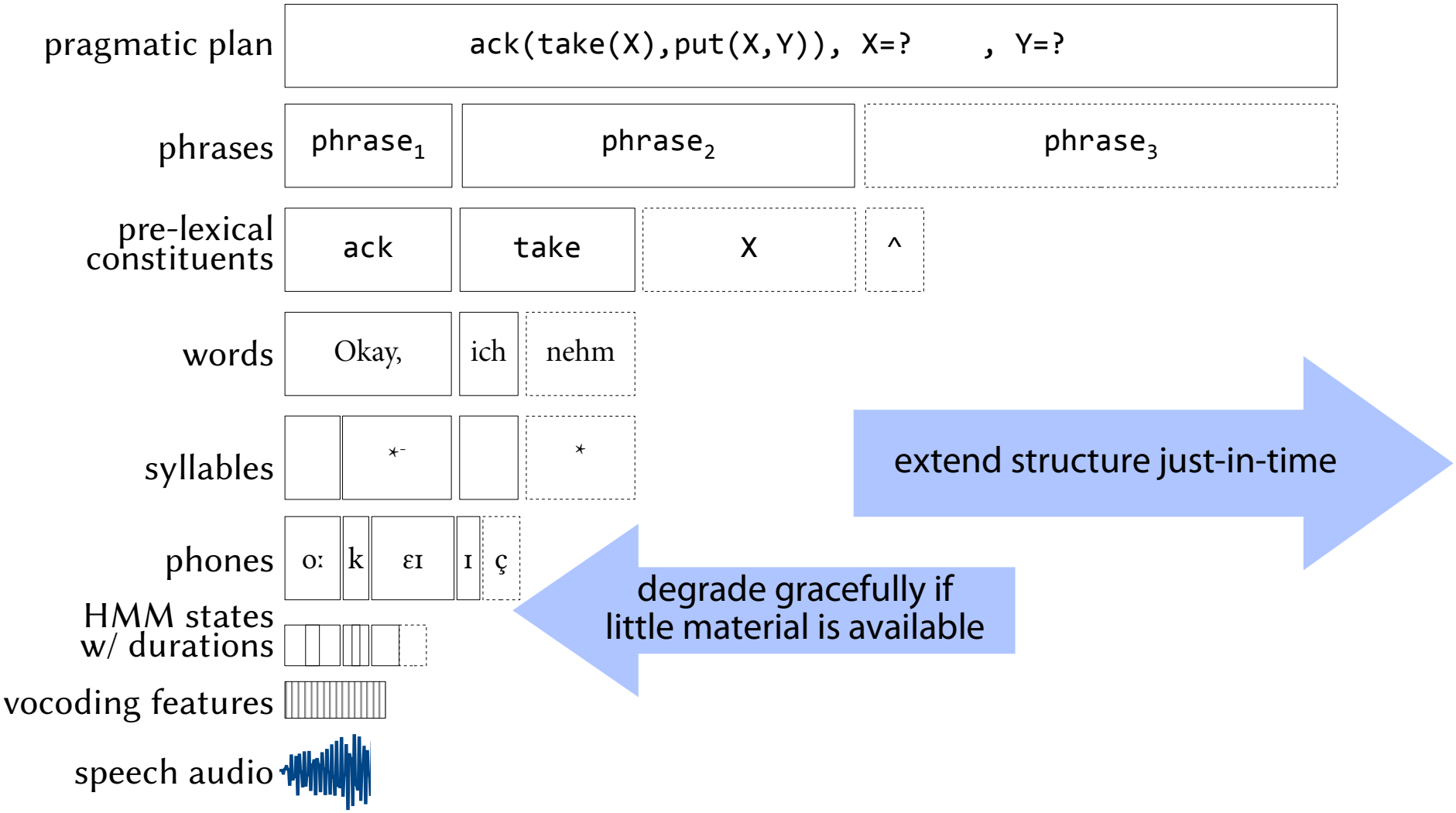
Baumann and Schlangen (2012a,b)
Baumann and Schlangen (2013a,b)
Baumann (2014a,b)

Incremental Speech Generation and Synthesis (HMM-based)



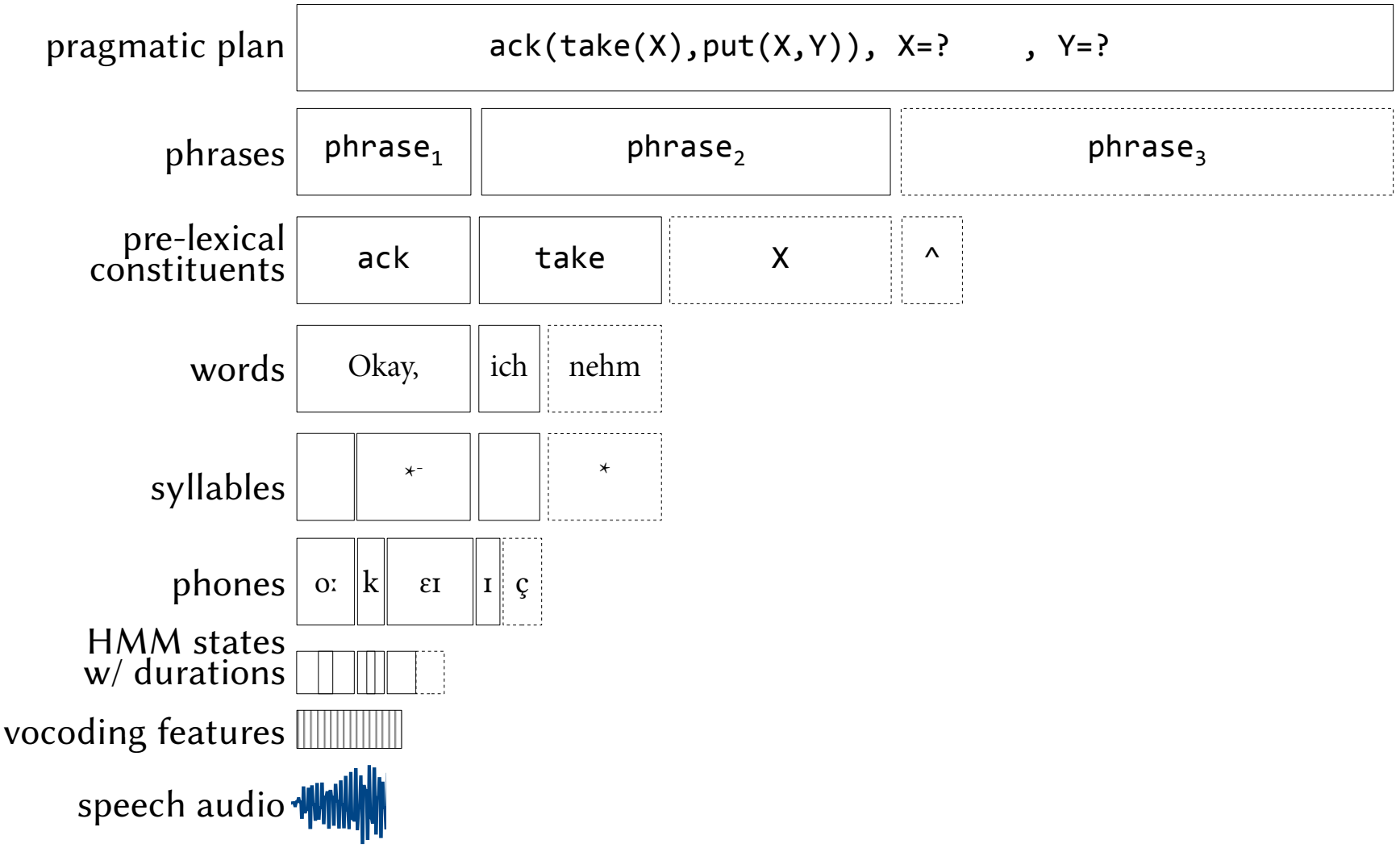
Baumann and Schlangen (2012a,b)
Baumann and Schlangen (2013a,b)
Baumann (2014a,b)

Incremental Speech Generation and Synthesis (HMM-based)

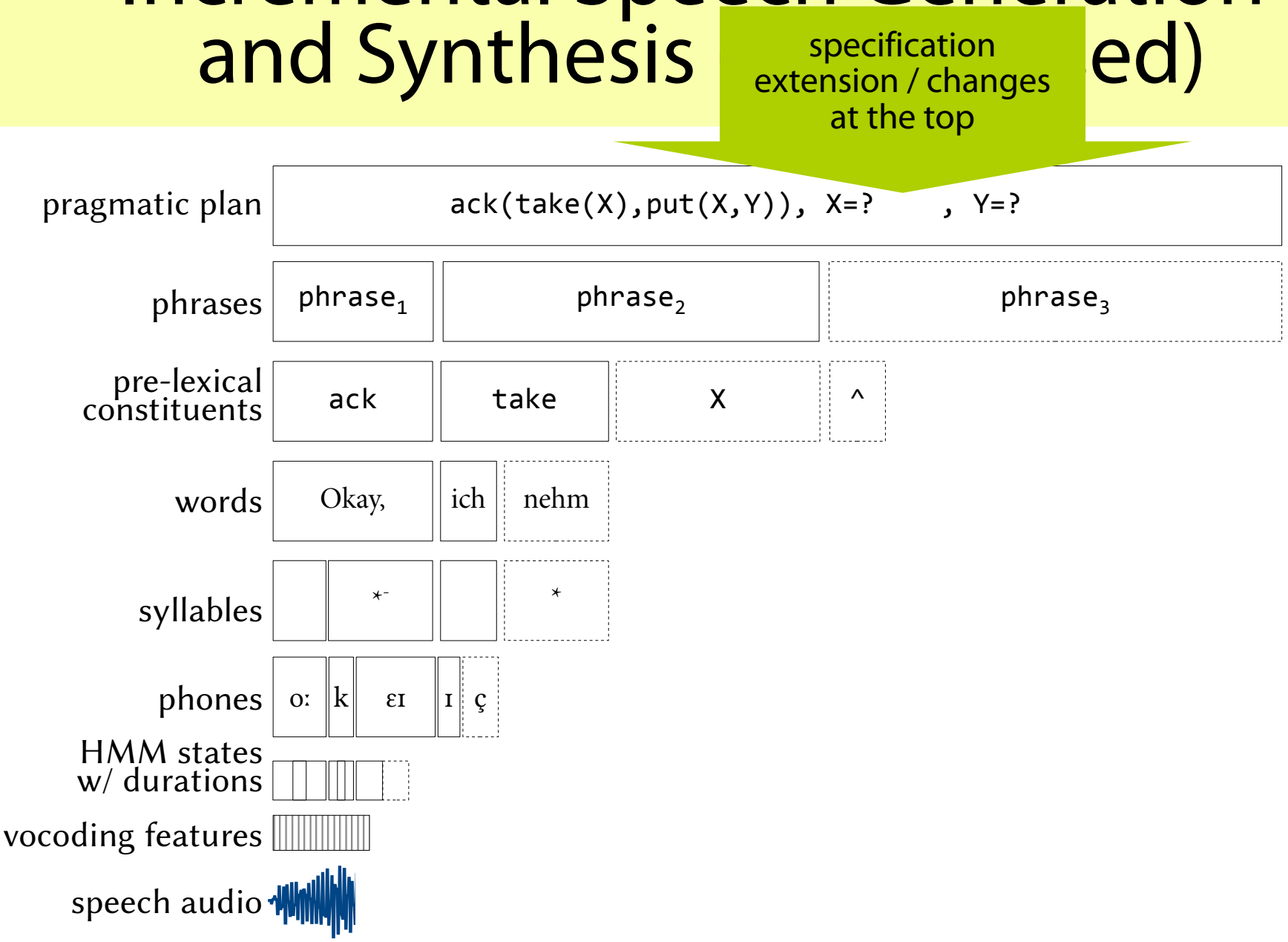


Baumann and Schlangen (2012a,b)
Baumann and Schlangen (2013a,b)
Baumann (2014a,b)

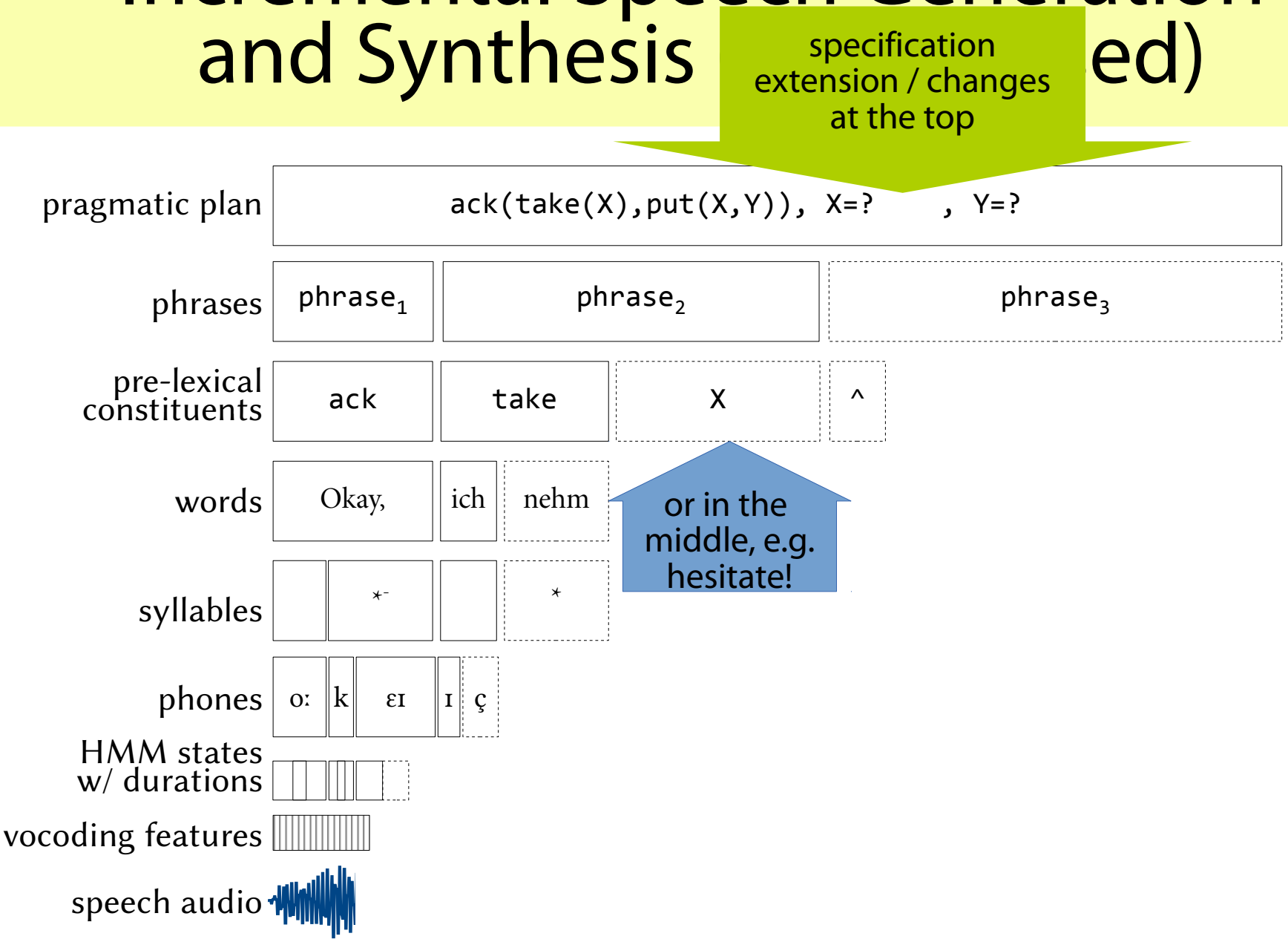
Incremental Speech Generation and Synthesis (HMM-based)



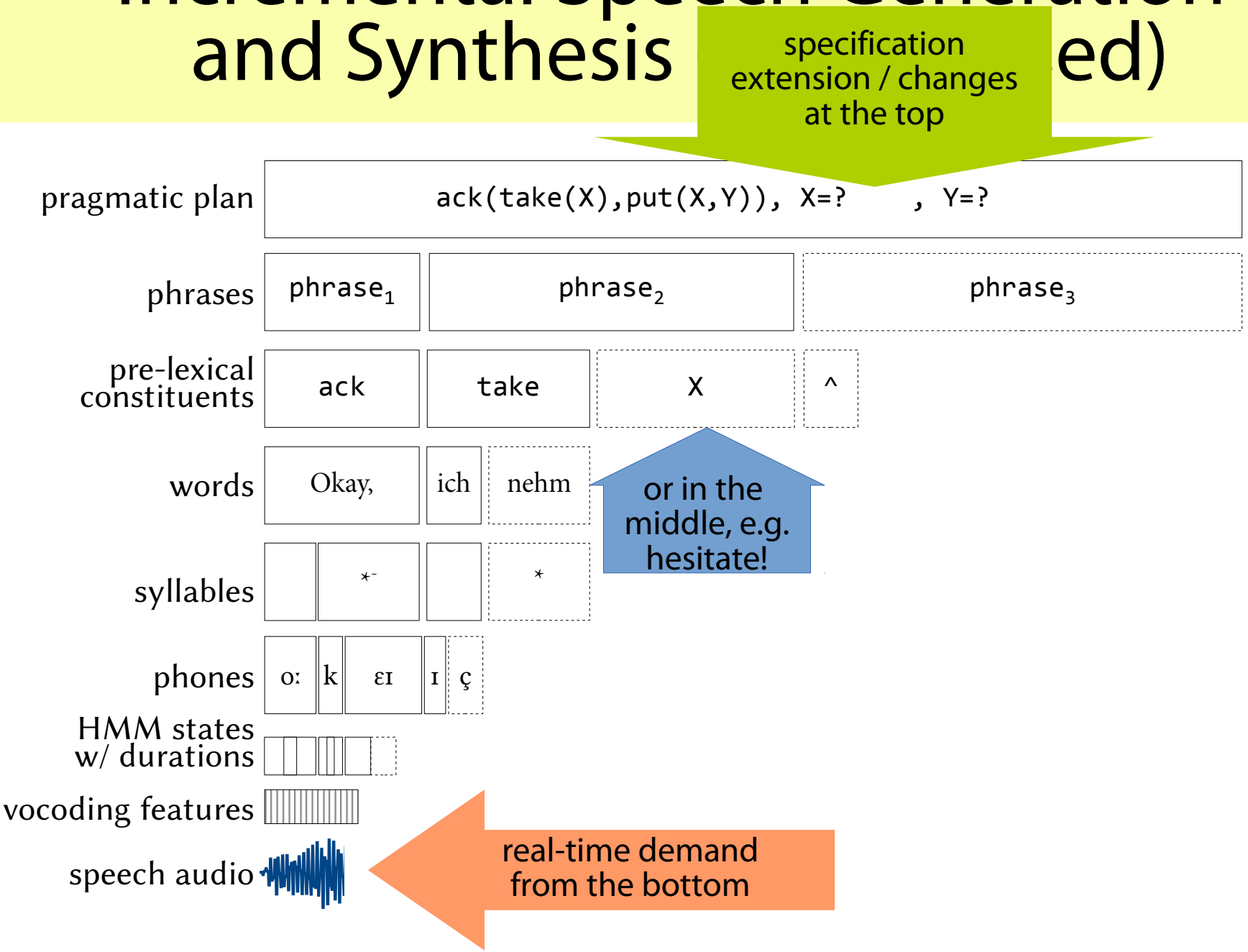
Incremental Speech Generation and Synthesis ed)



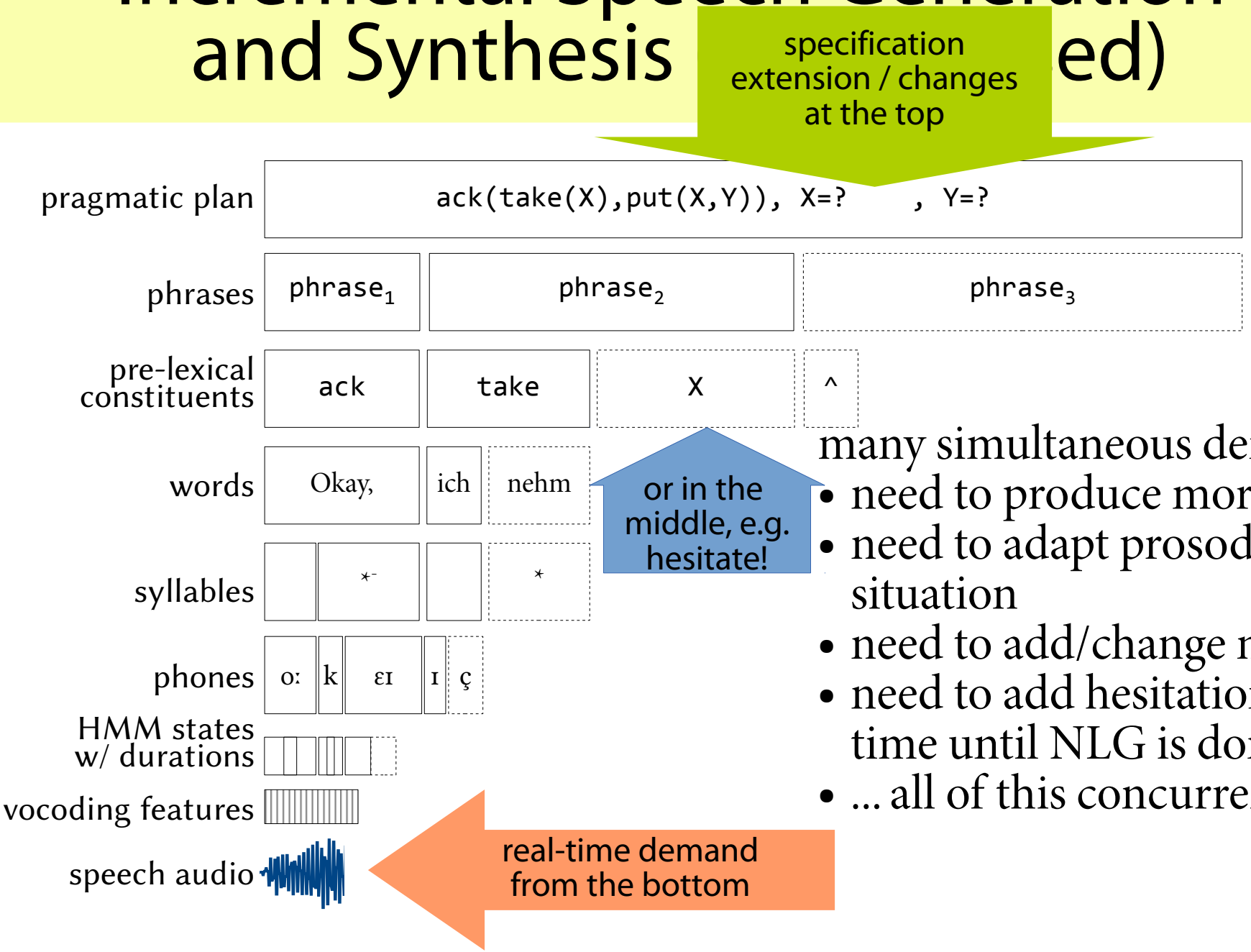
Incremental Speech Generation and Synthesis ed)



Incremental Speech Generation and Synthesis ed)



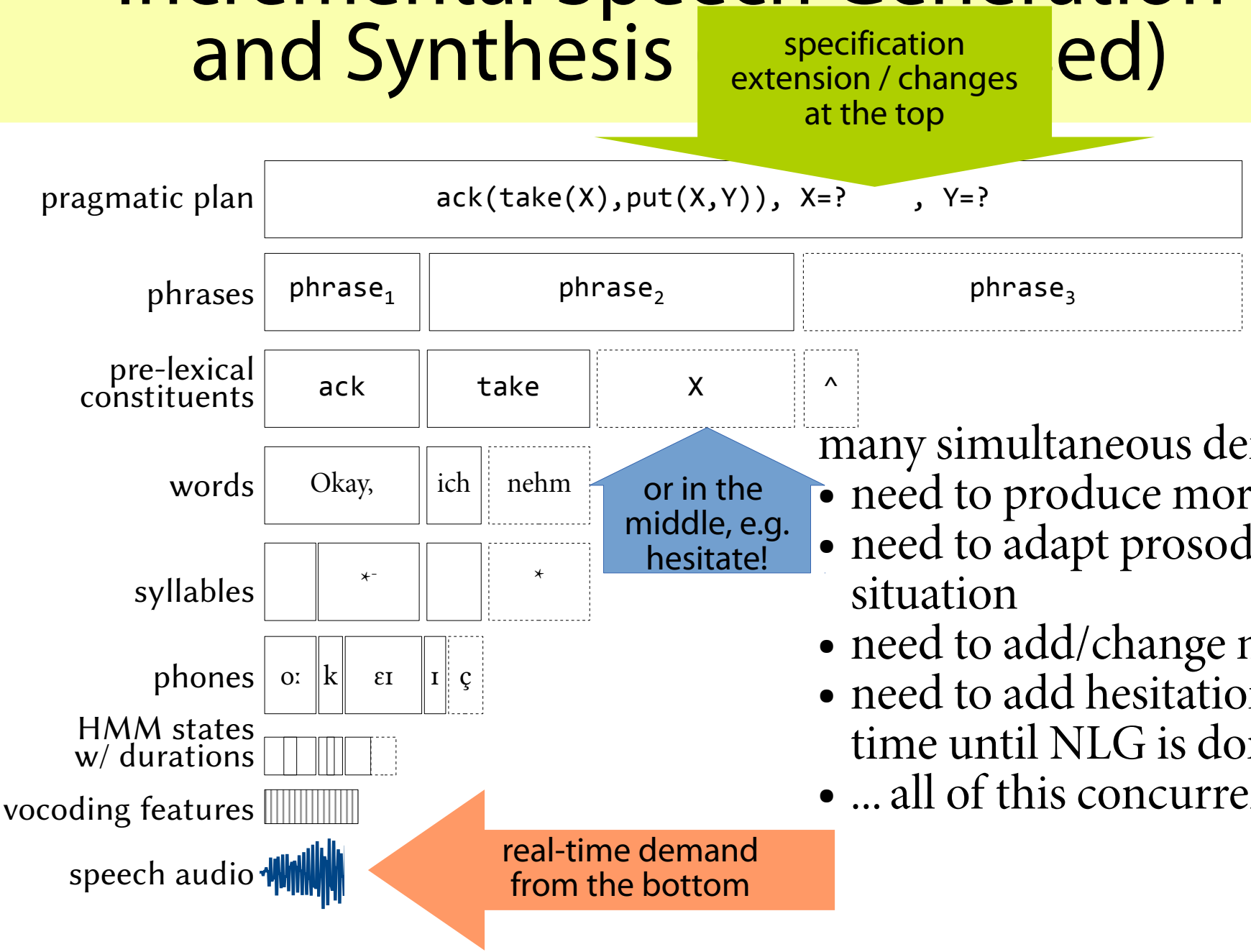
Incremental Speech Generation and Synthesis ed)



many simultaneous demands:

- need to produce more speech
- need to adapt prosody to situation
- need to add/change material
- need to add hesitation to span time until NLG is done
- ... all of this concurrently

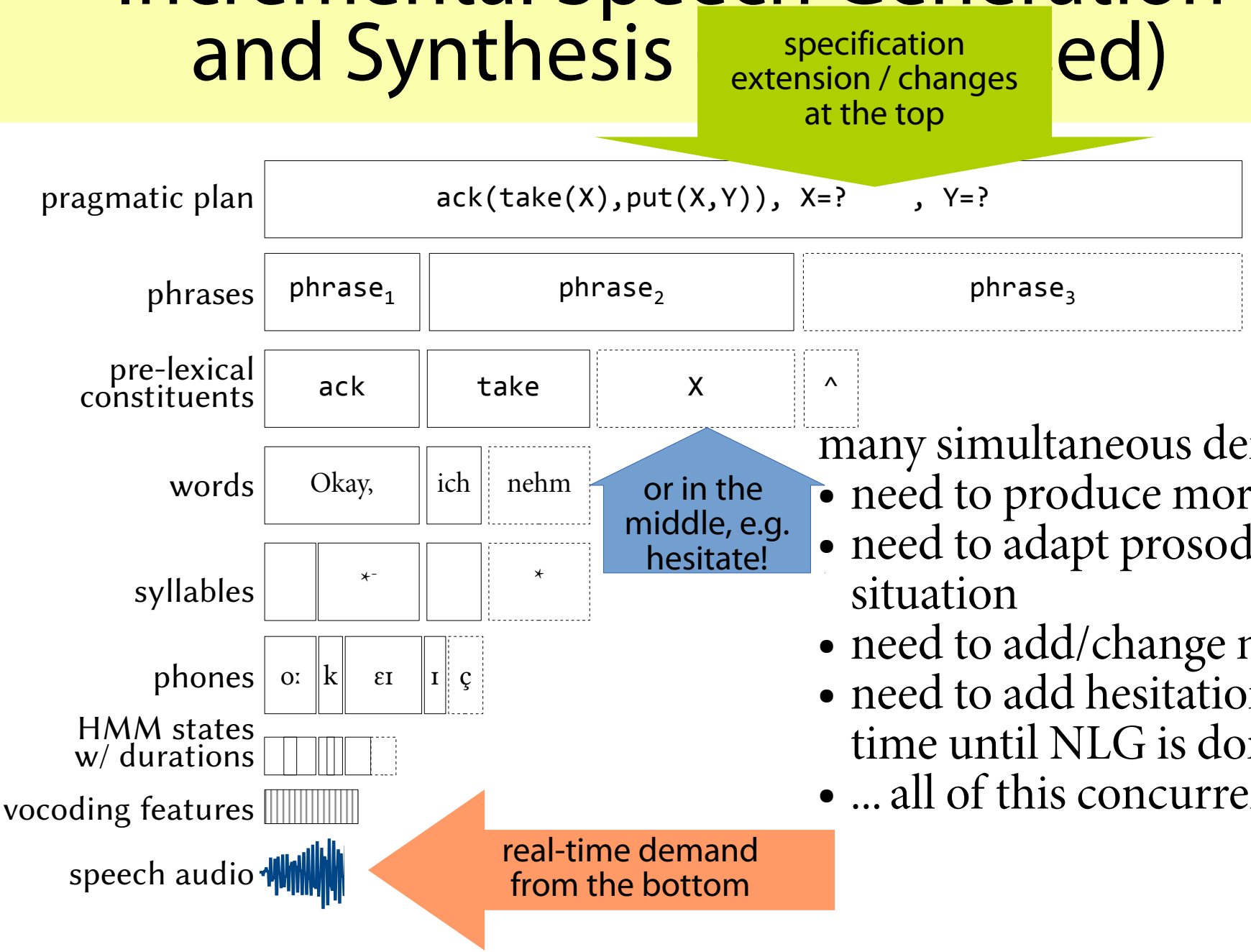
Incremental Speech Generation and Synthesis ed)



many simultaneous demands:

- need to produce more speech
- need to adapt prosody to situation
- need to add/change material
- need to add hesitation to span time until NLG is done
- ... all of this concurrently

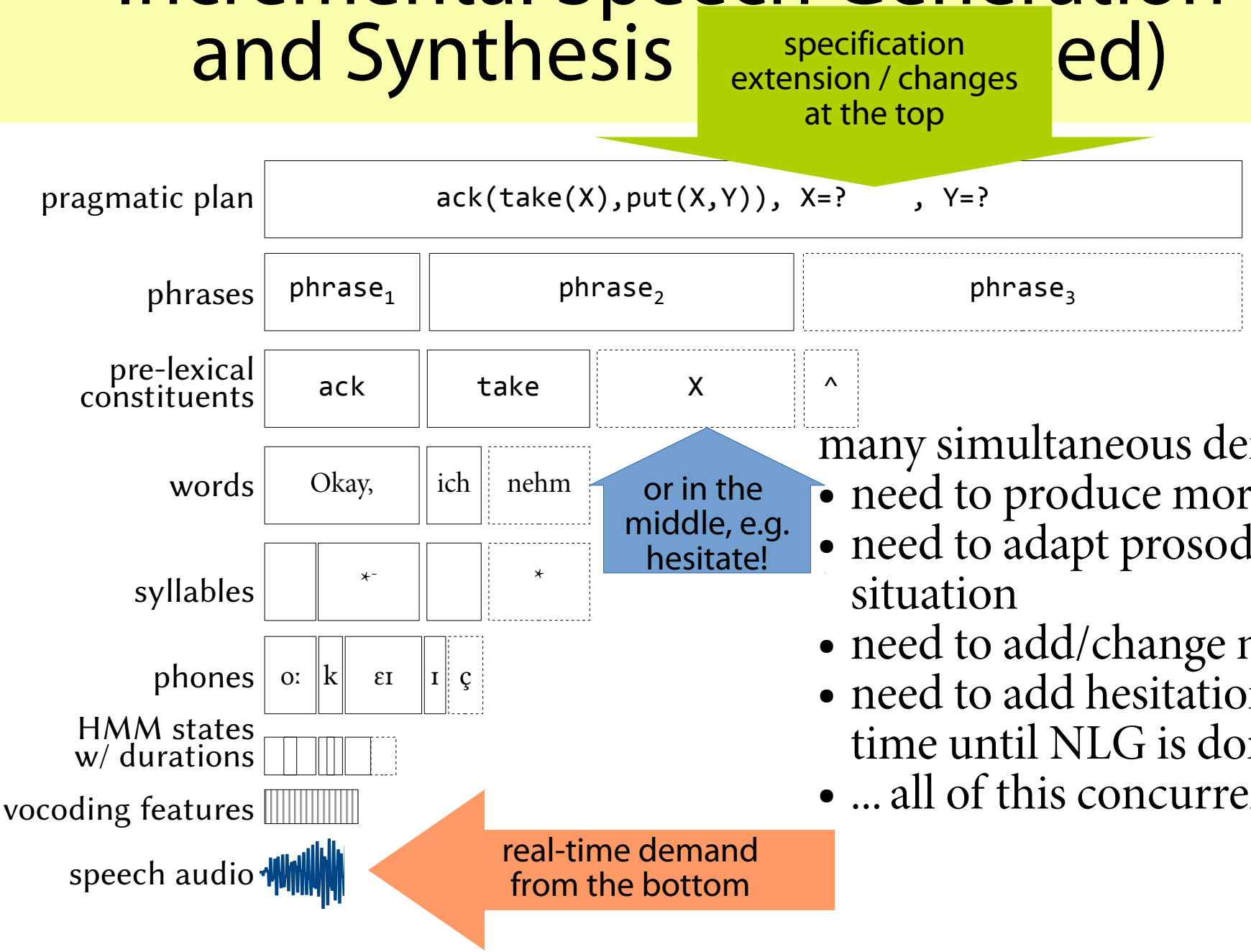
Incremental Speech Generation and Synthesis ed)



many simultaneous demands:

- need to produce more speech
- need to adapt prosody to situation
- need to add/change material
- need to add hesitation to span time until NLG is done
- ... all of this concurrently

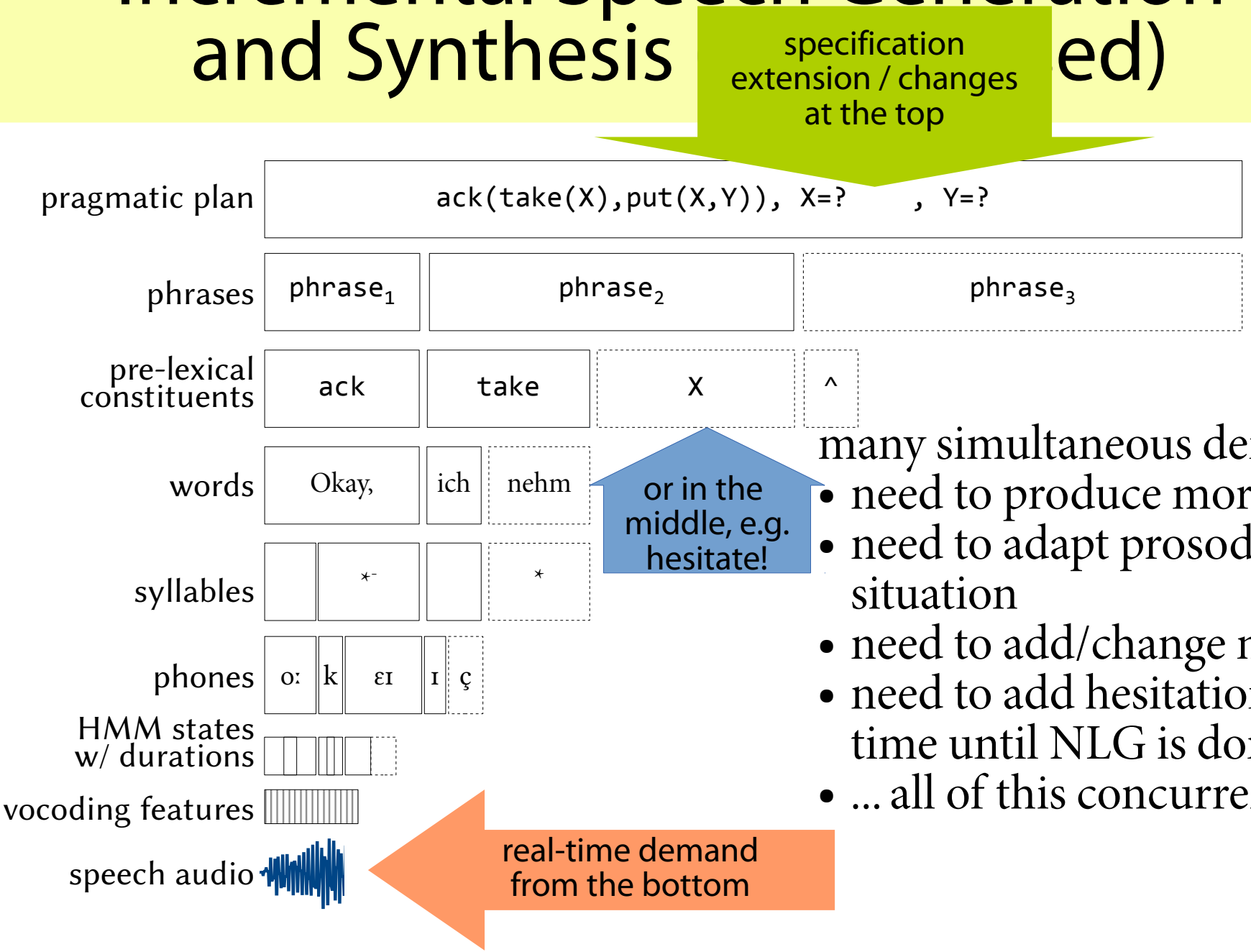
Incremental Speech Generation and Synthesis ed)



many simultaneous demands:

- need to produce more speech
- need to adapt prosody to situation
- need to add/change material
- need to add hesitation to span time until NLG is done
- ... all of this concurrently

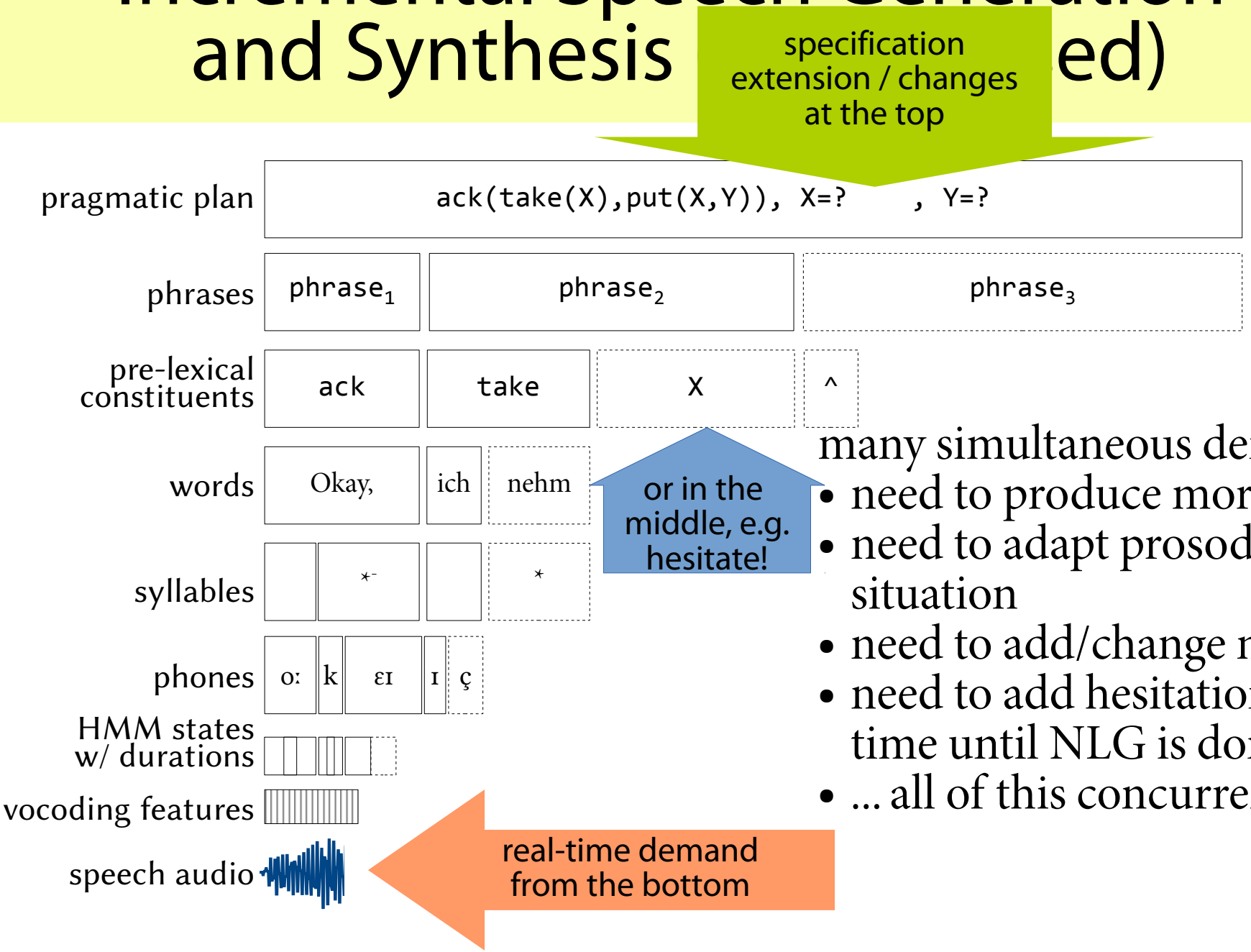
Incremental Speech Generation and Synthesis ed)



many simultaneous demands:

- need to produce more speech
- need to adapt prosody to situation
- need to add/change material
- need to add hesitation to span time until NLG is done
- ... all of this concurrently

Incremental Speech Generation and Synthesis ed)



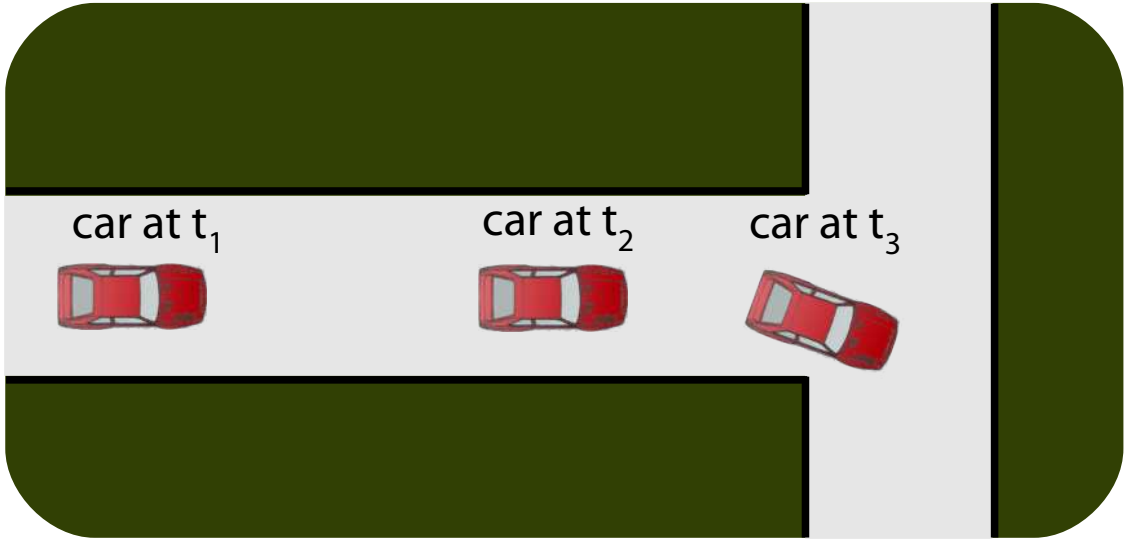
many simultaneous demands:

- need to produce more speech
- need to adapt prosody to situation
- need to add/change material
- need to add hesitation to span time until NLG is done
- ... all of this concurrently

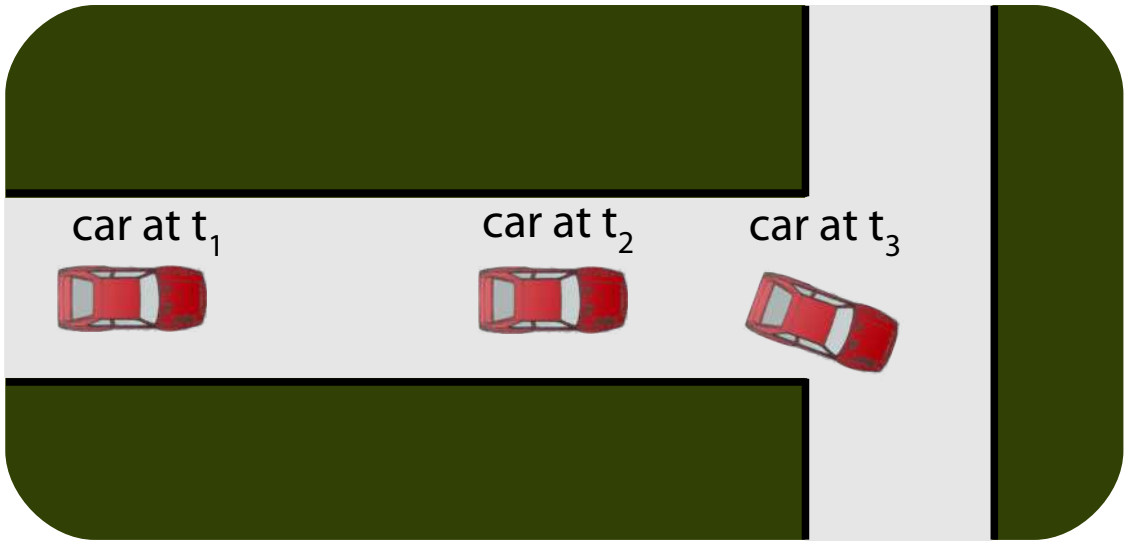

Breaking
NEWS

Acting in Dynamic Environments

- dynamic environment changes quickly
  - rate of notable events is high – too high to generate one descriptive utterance per event
- events need to be combined into complex utterances



Acting Incrementally



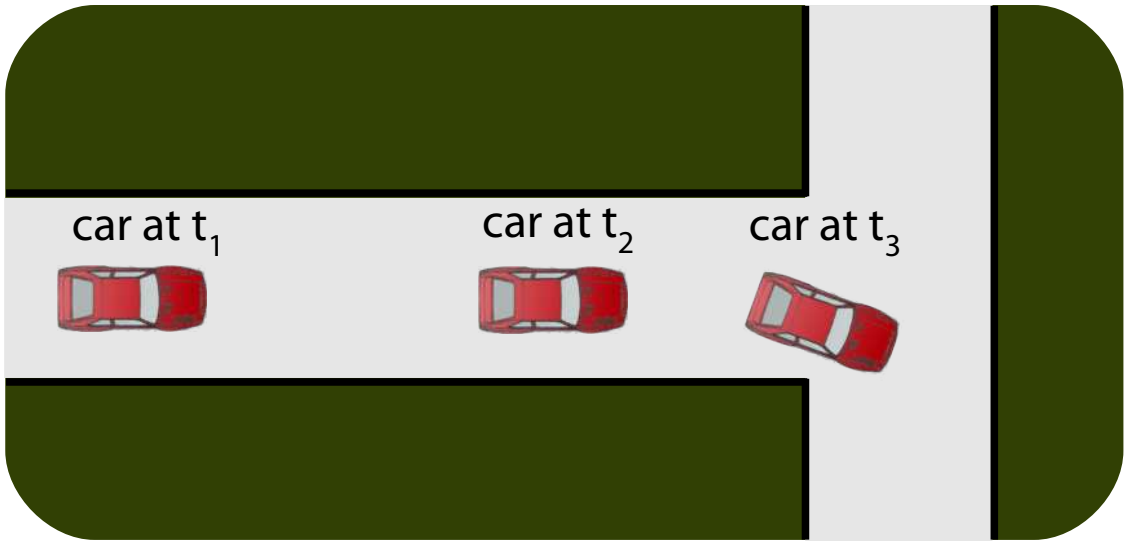
- at t_1: car drives along street
- at t_2: the car is *likely* to turn
- at t_3: car is turning right

"The car drives along X-street.

...X-street, and then turns...

...right into Y-street."

Acting Incrementally



- at t_1: car drives along street
- at t_2: the car is *likely* to turn
- at t_3: car is turning right

"The car drives along X-street.

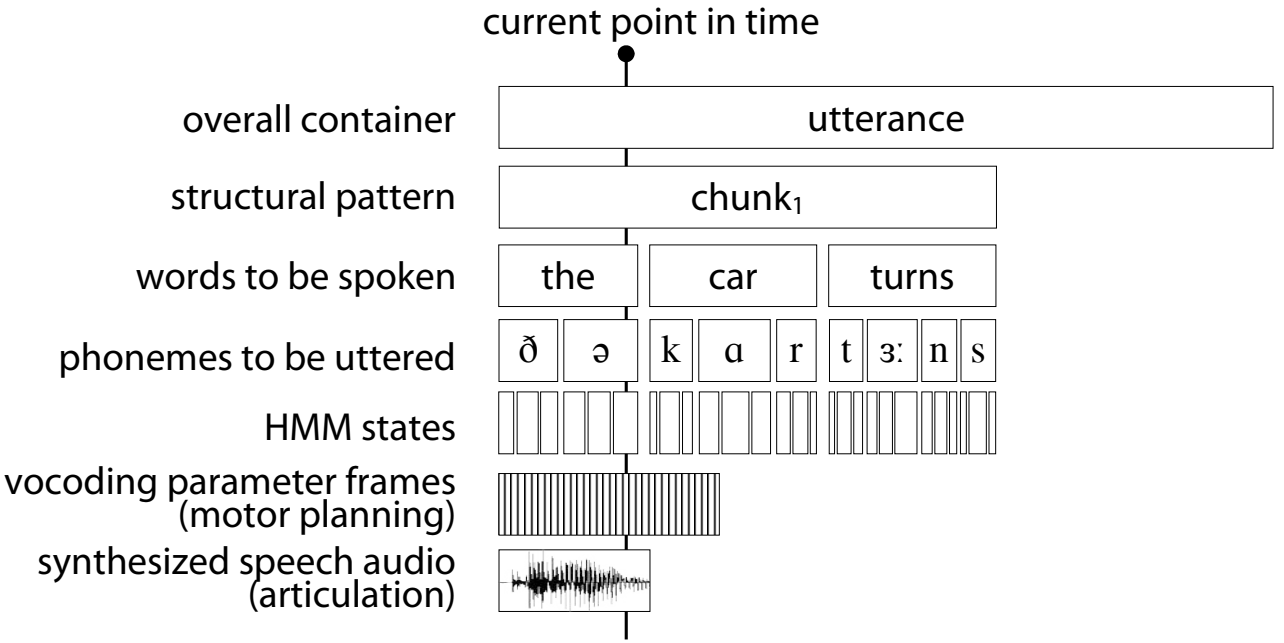
...X-street, and then turns...

...right into Y-street."

prosody must be adapted!

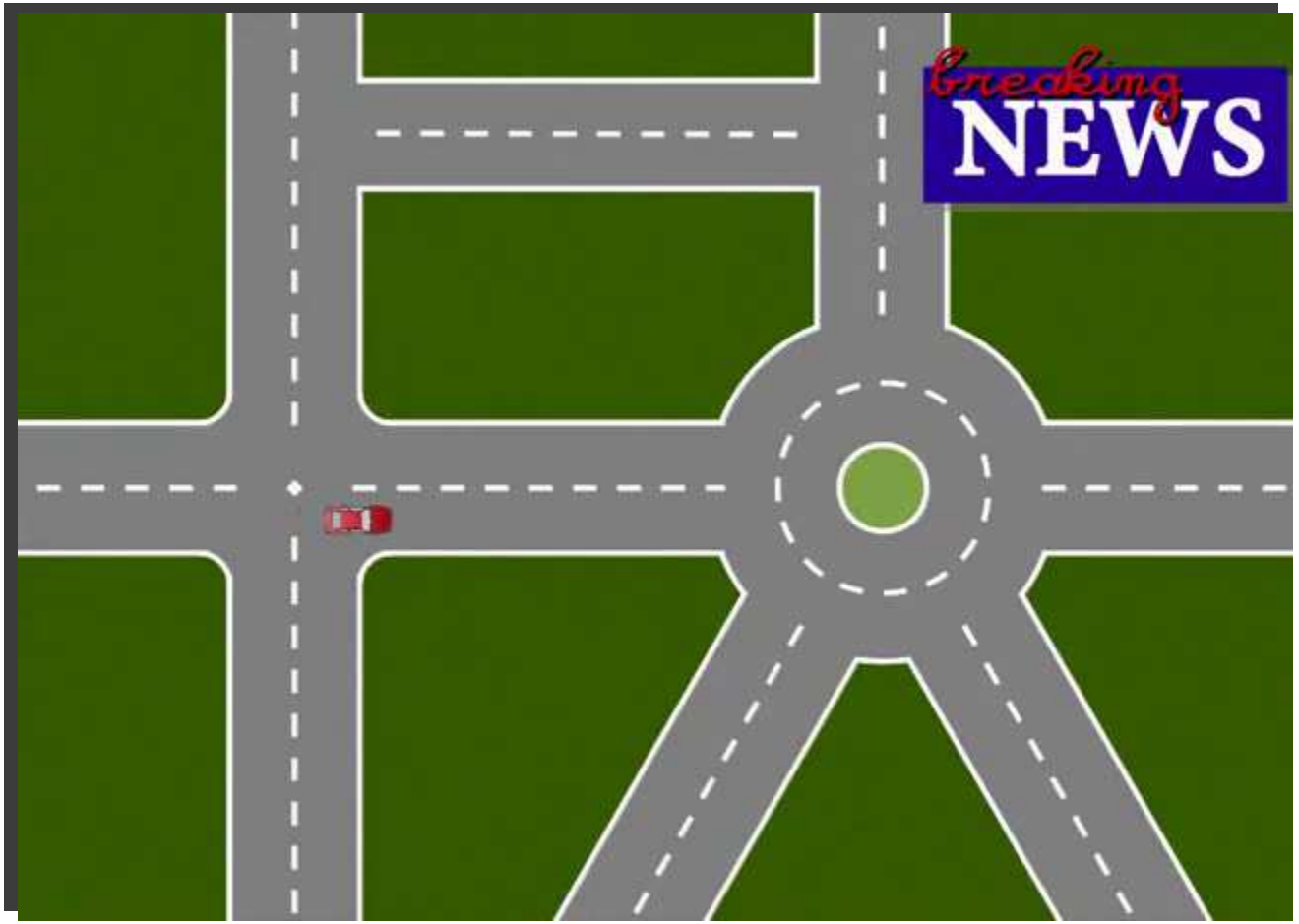
Acting Incrementally: Challenges

- *piece-meal input*: phrases or individual words
  - conventional speech synthesis assumes full utterances
- phrases have to be *connected prosodically*
- processing should occur *just-in-time*



given that incremental speech synthesis
measurable degrades prosodic parameters –
→ does this degradation matter to listeners?

Standard behaviour



Incremental behaviour (taking expectations into account)

Experiment

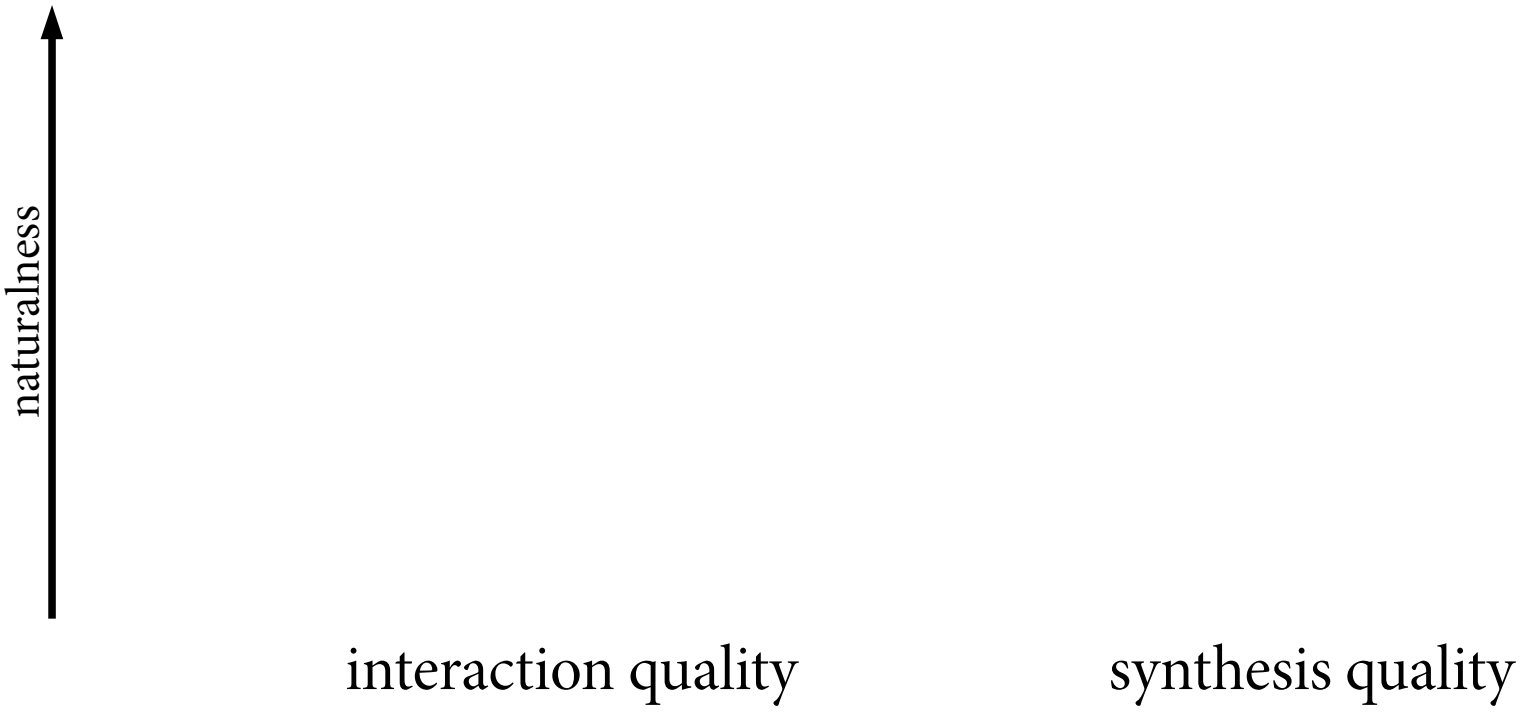
- incremental system vs. baseline system
- 9 settings in the CarChase domain
- 9 subjects were asked to rate (5-point Likert)
  - naturalness of verbalization (to capture interactional adequacy)
  - naturalness of *pronunciation* (to capture synthesis quality)
- results in 81 paired samples

- incremental processing implemented in InproTK, using speech synthesis technology from MaryTTS

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

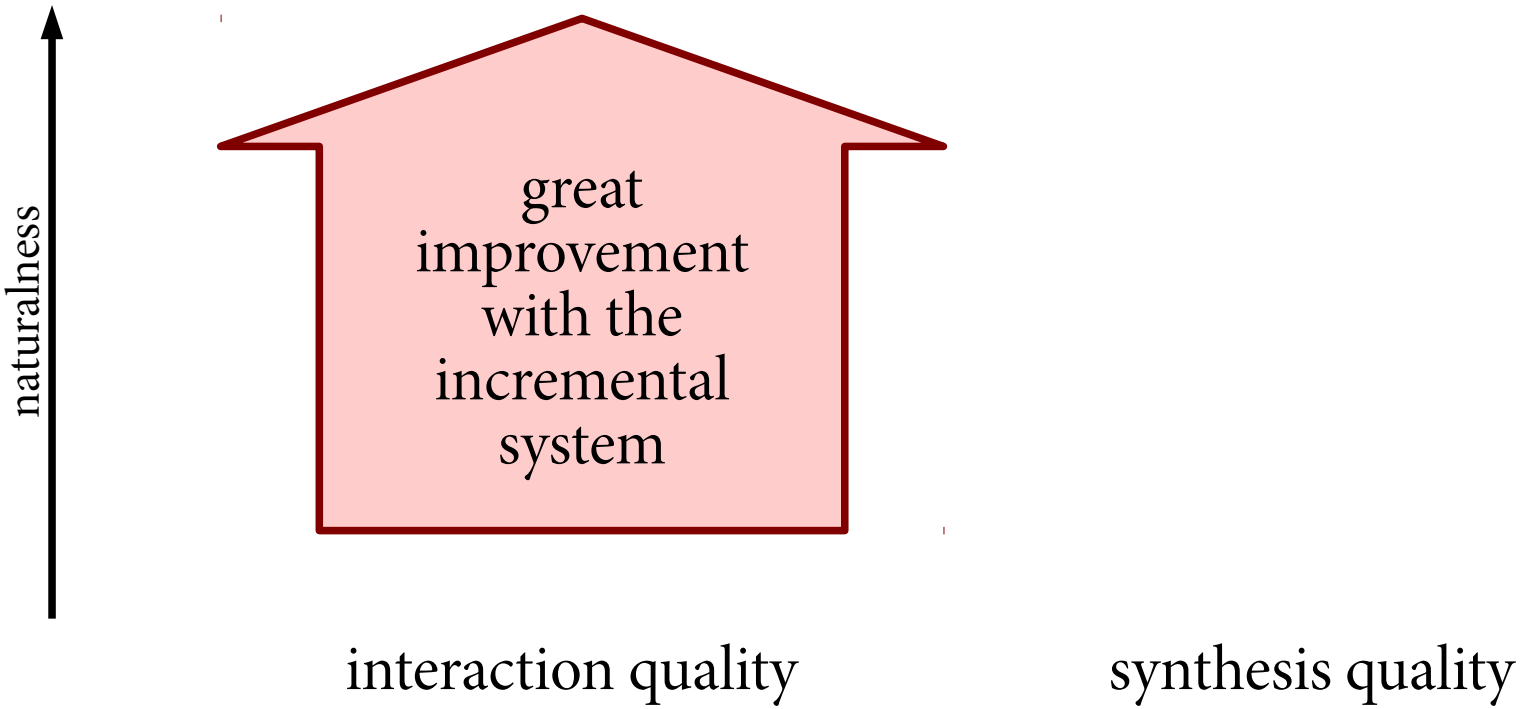
Expected results

- we were hoping for a good trade-off:



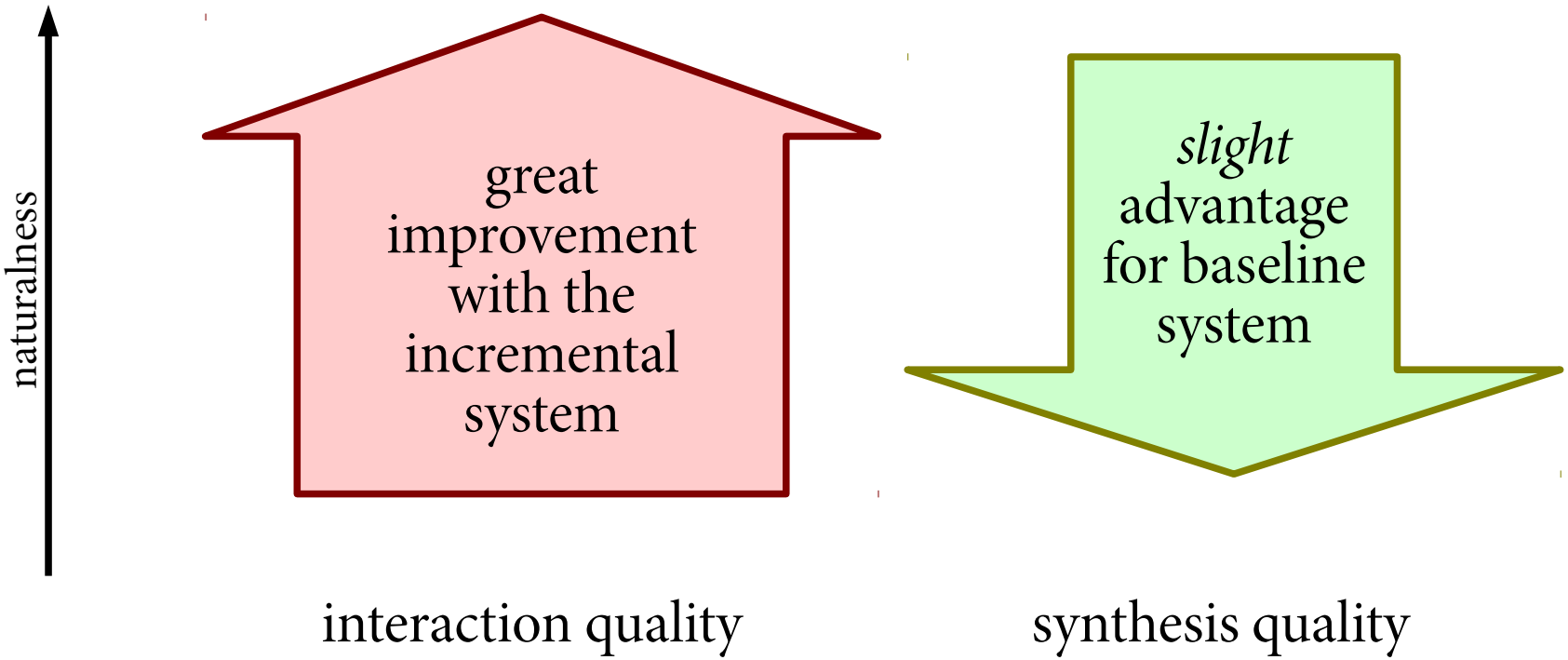
Expected results

- we were hoping for a good trade-off:



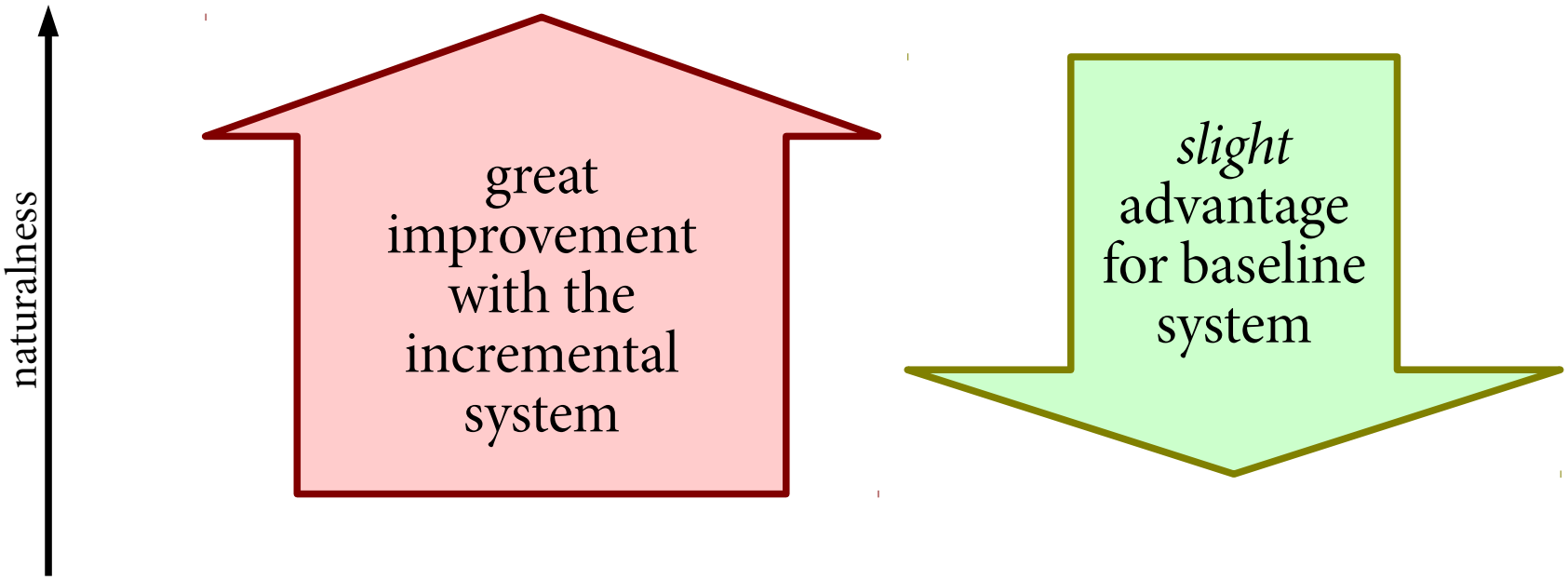
Expected results

- we were hoping for a good trade-off:



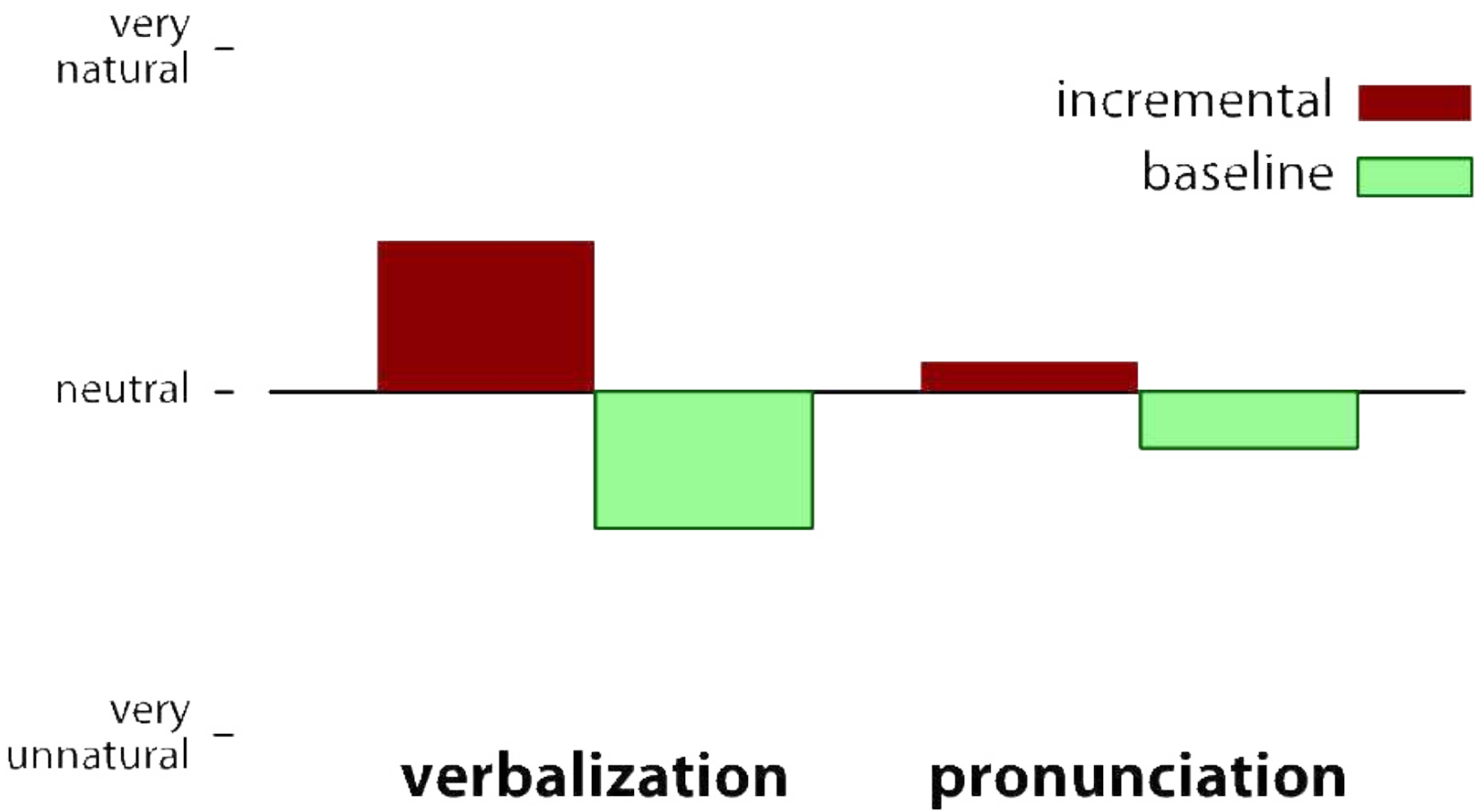
Expected results

- we were hoping for a good trade-off:

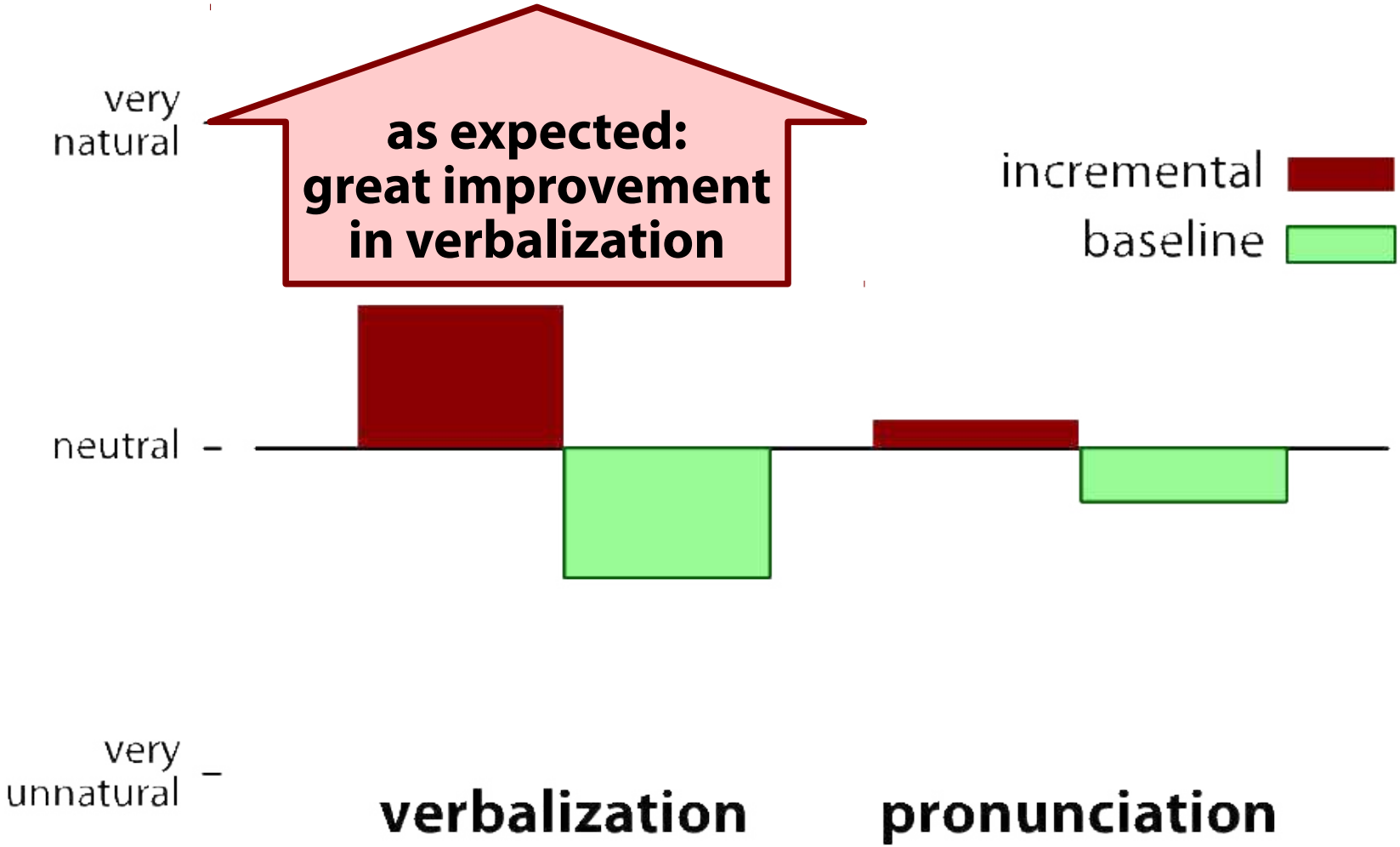


→ write paper: „Trade-off between incrementality of behaviour and speech synthesis quality“

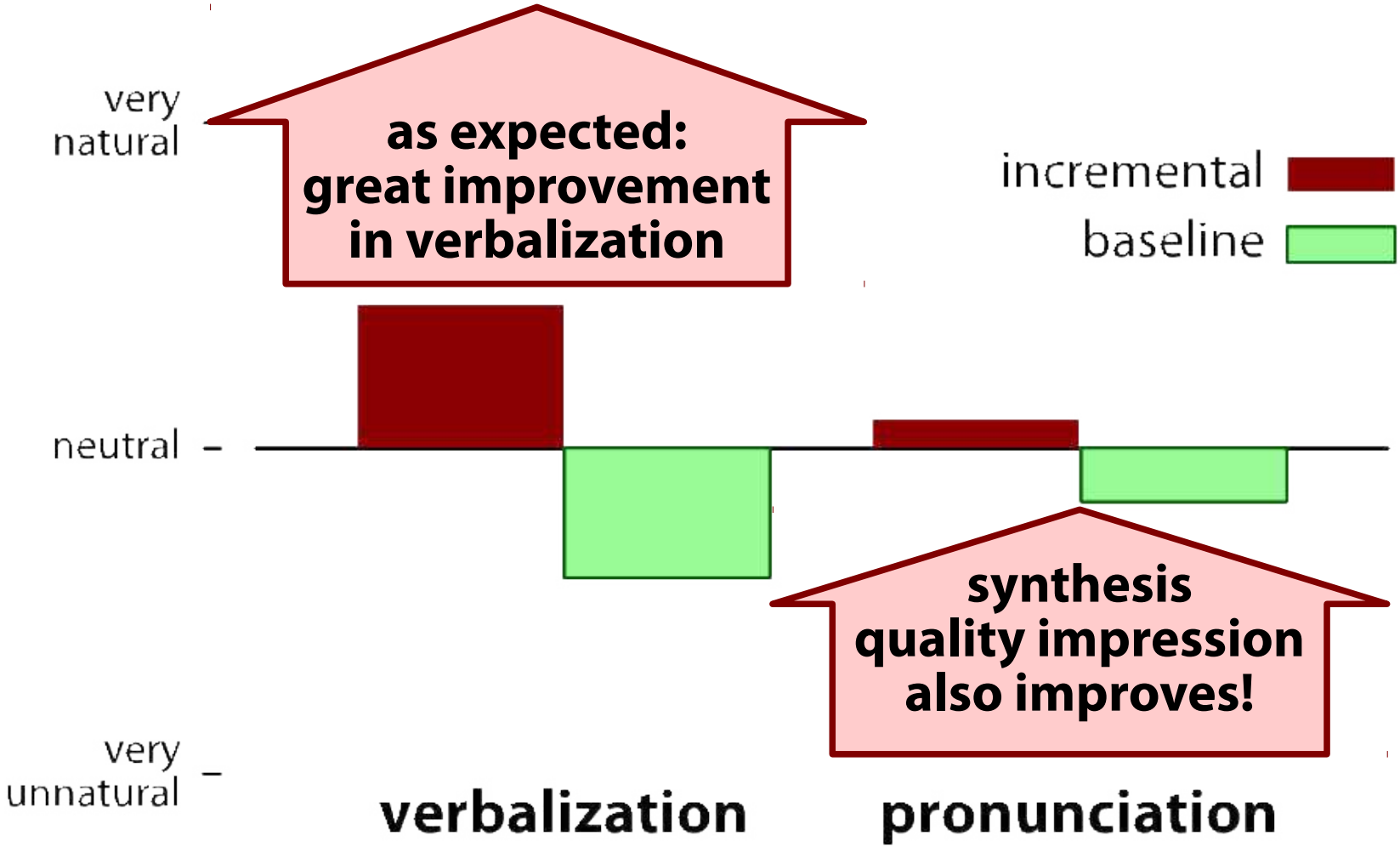
Actual results



Actual results



Actual results



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

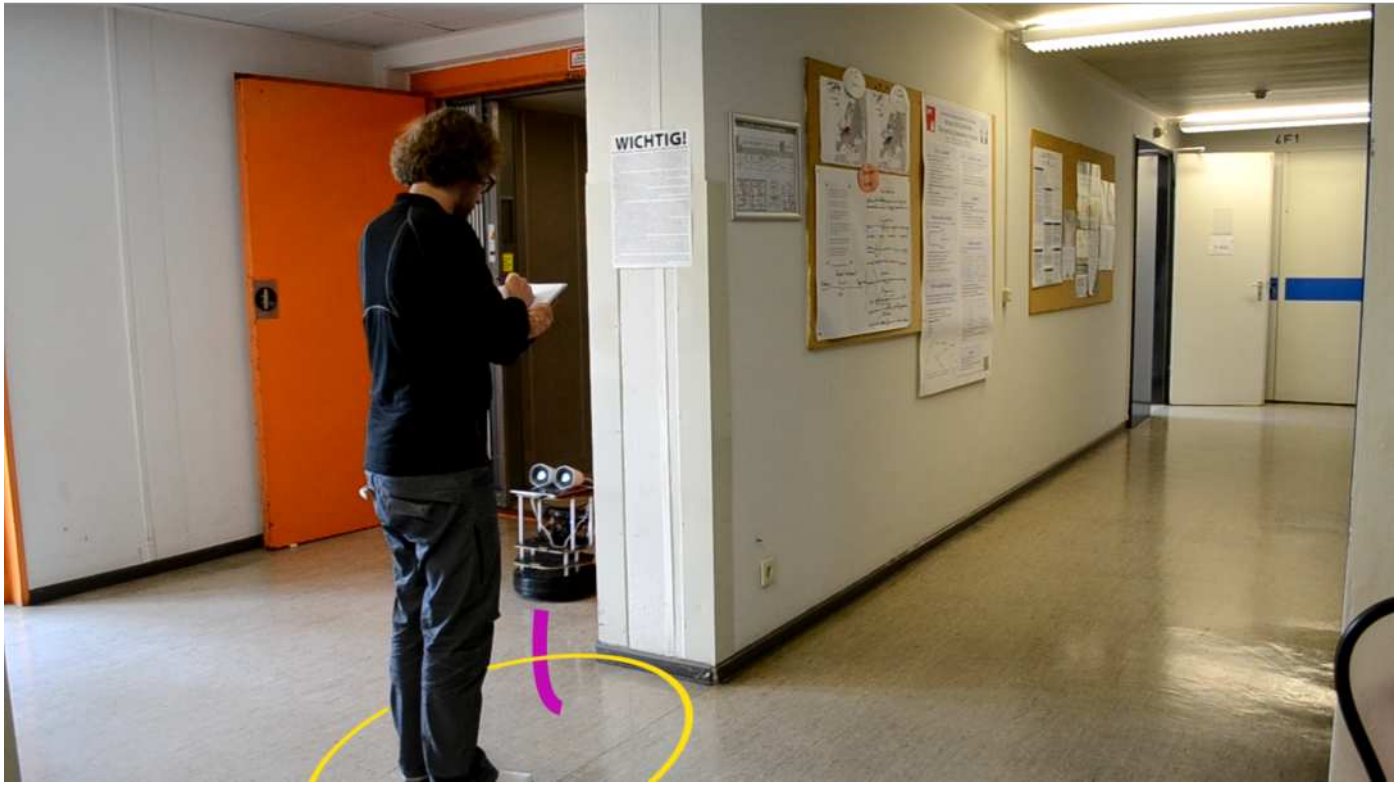
Conclusions: Incremental speech output

- adequate verbalization / wording in a given context
  - may be more important than synthesis quality
  - may even lead to better synthesis quality ratings!
  - despite somewhat reduced quality

➔ you need to find out what really matters for the users of your application!

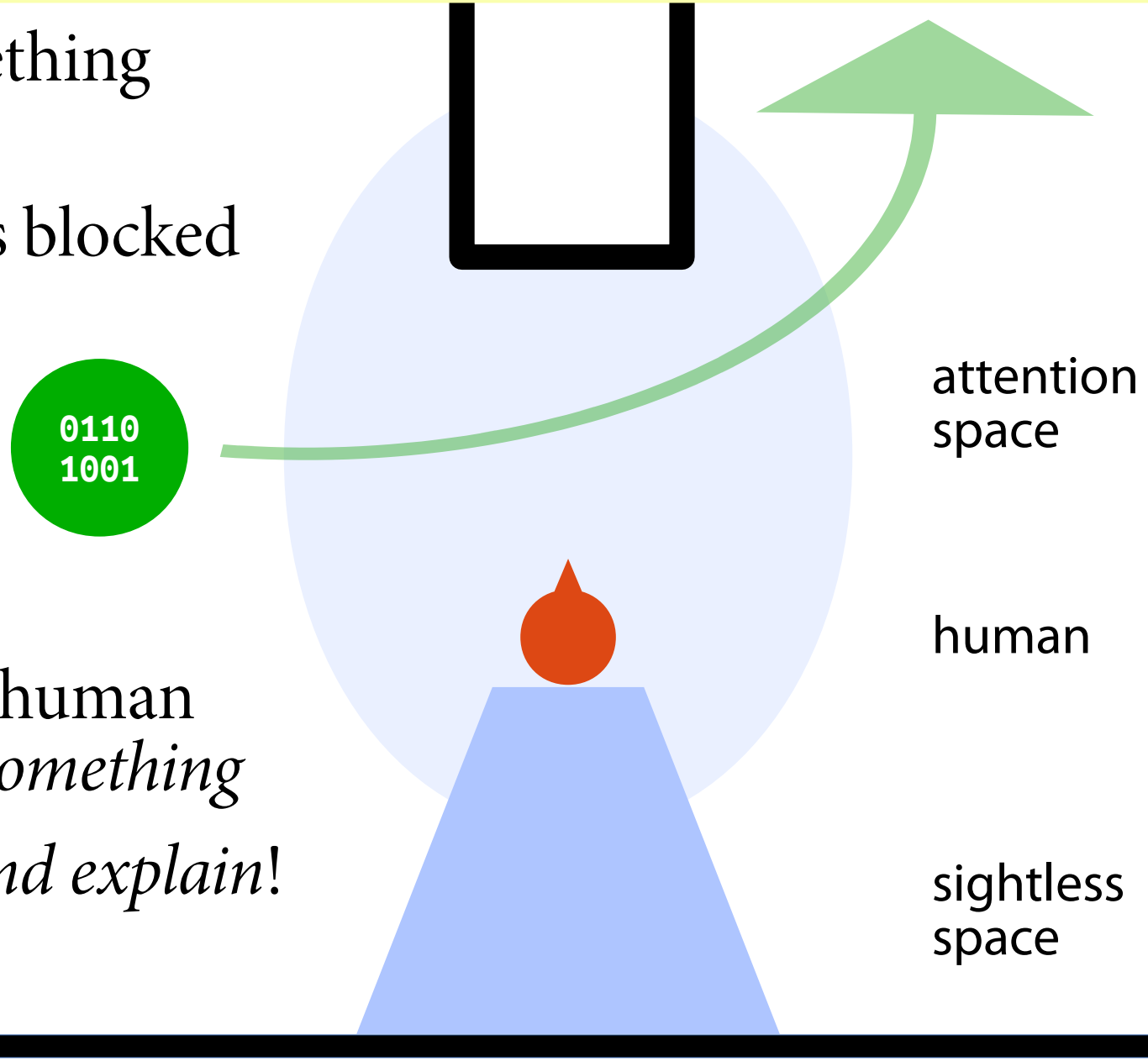
Incremental speech output in social robot navigation

- personal space intrusion:
  dispreferred but sometimes necessary



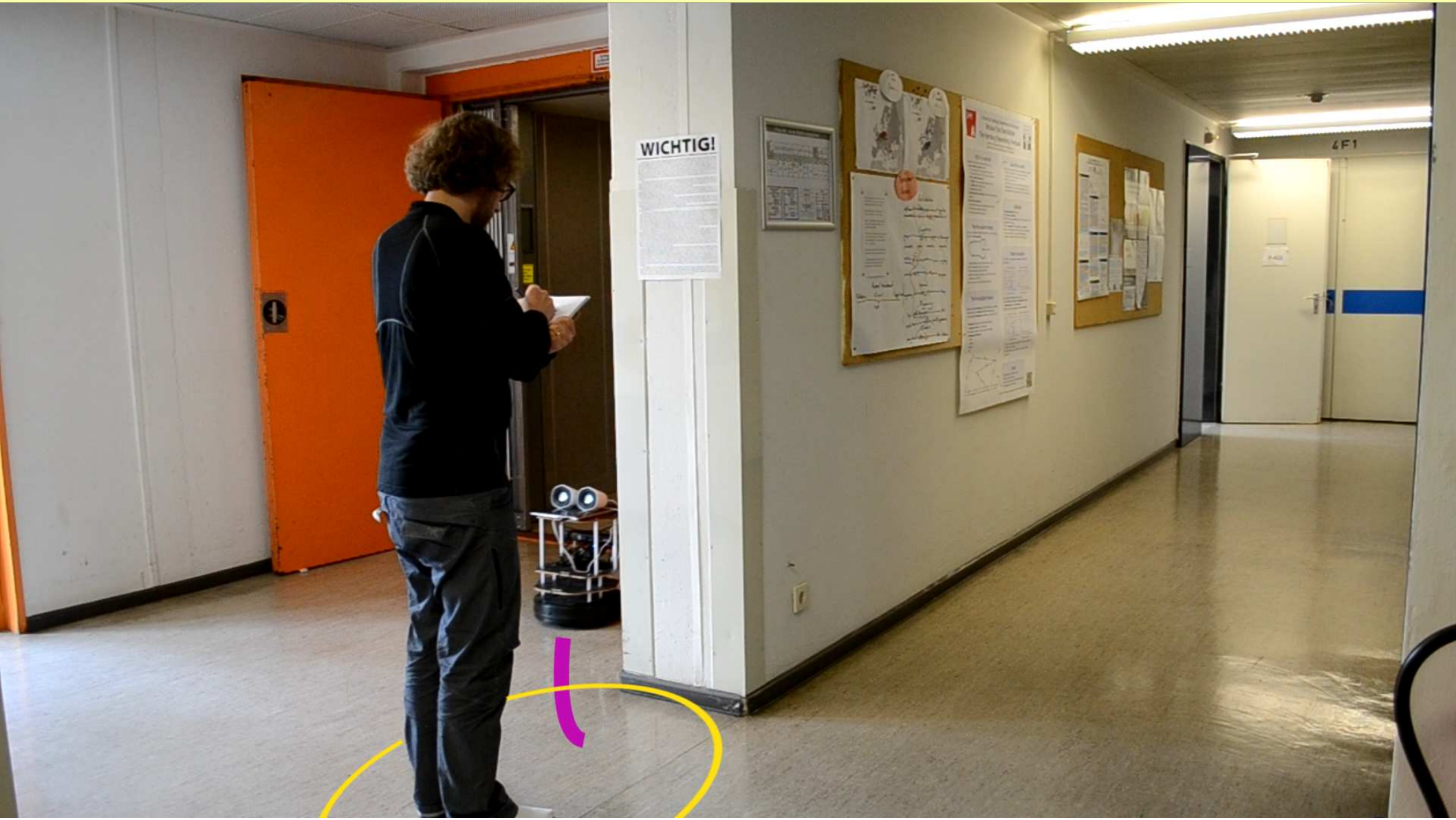
Personal Space Intrusion

- human reads something in front of her
- space behind her is blocked



- impossible to pass human without violating *something*
- at least *apologize and explain*!

Example



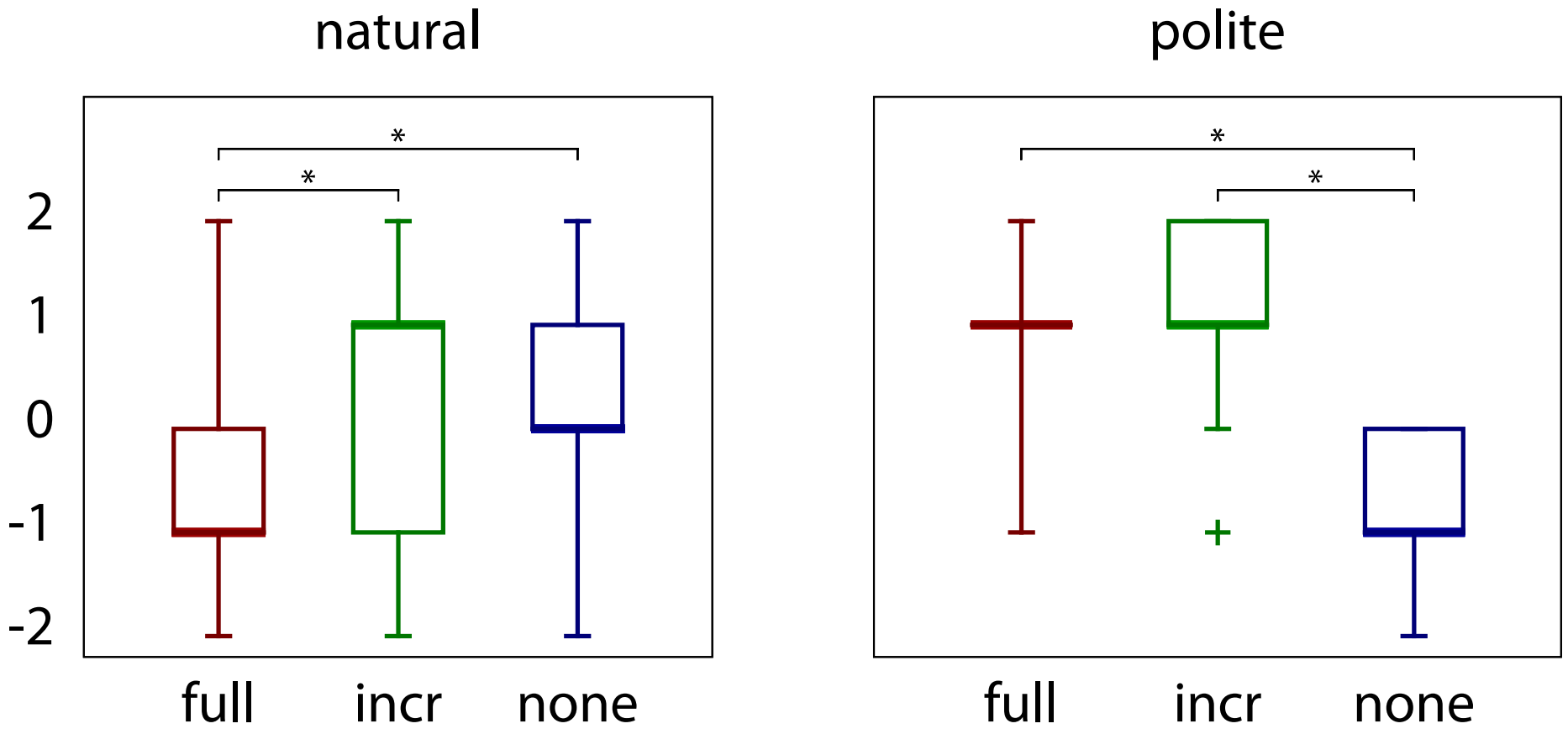

WICHTIG!

Controlled study

- simulated robot environment
- different robot speed/human movement conditions
- three speaking conditions:
  - silent
  - apologize, explain and thank when entering the personal space
    → what if we only touch the personal space briefly?
  - skip forward to „thank you“ when we exit the personal space

- 13 students, 12 videos, three questions:
  robot naturalness, politeness, route&speed

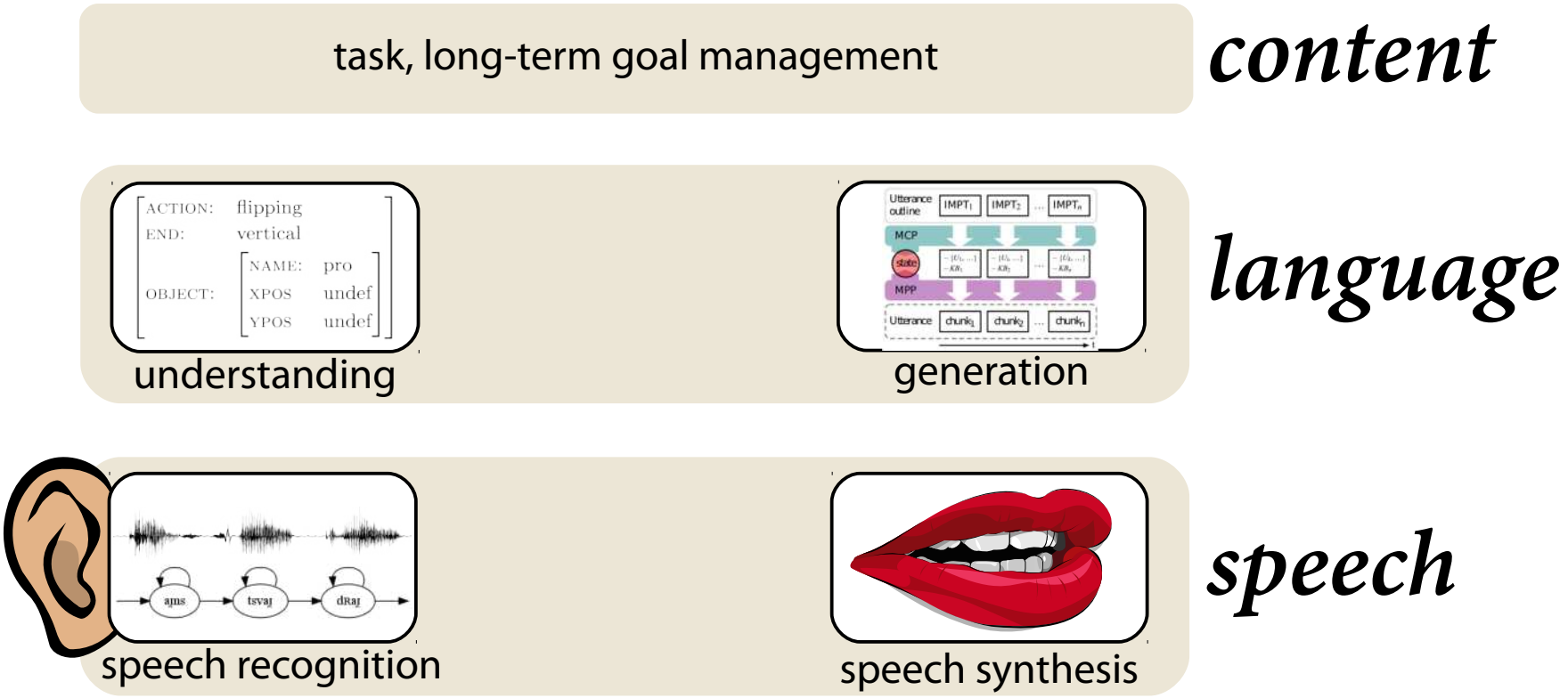
Results



- we find that the incremental speech production strategy is both natural and polite.

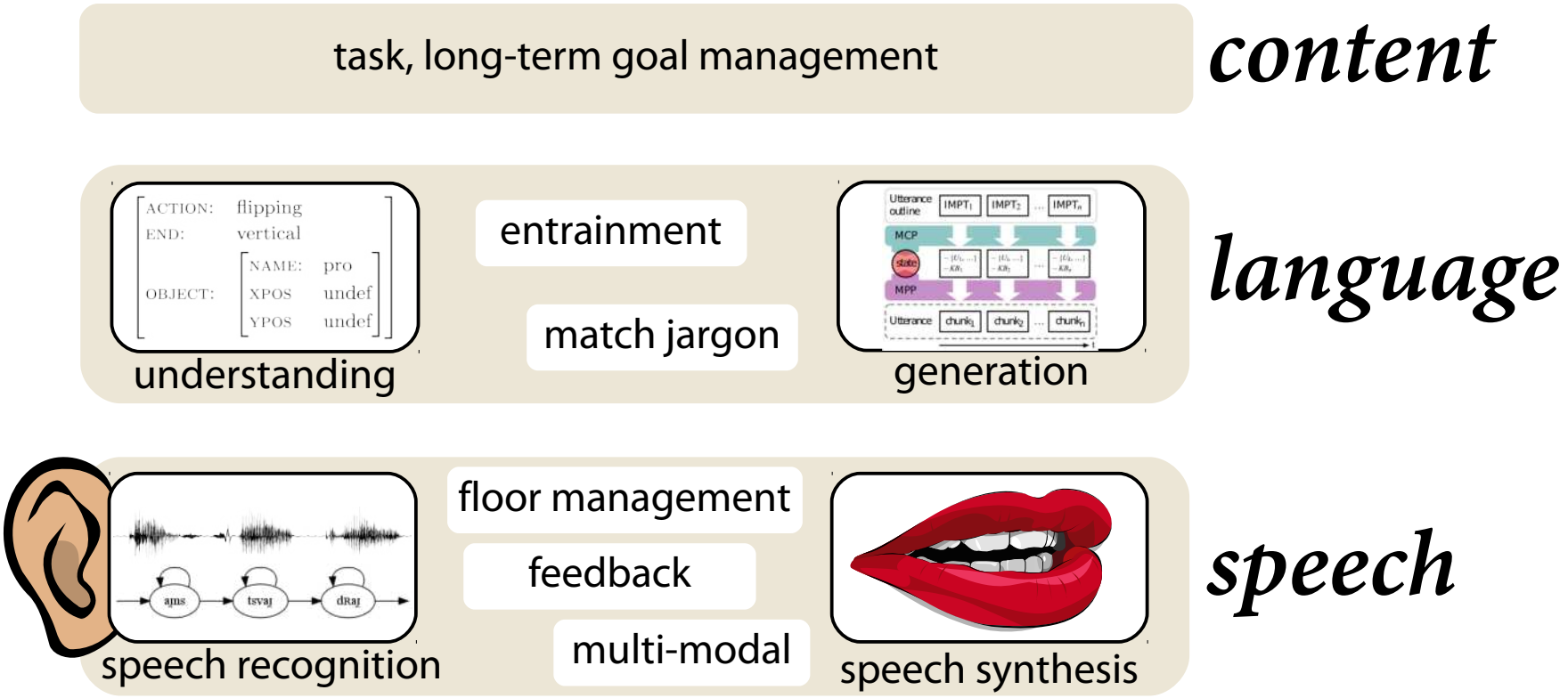
Langzeitperspektive auf Incrementalität und Integration

- "Stratifikation" der Abstraktionsebenen ermöglicht fortgeschrittene konversationale Fähigkeiten



Langzeitperspektive auf Incrementalität und Integration

- “Stratifikation” der Abstraktionsebenen ermöglicht fortgeschrittene konversationale Fähigkeiten

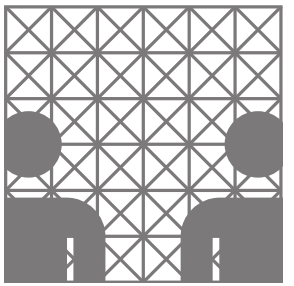


Zusammenfassung

- **Responsivität** und **Flexibilität** sind Schlüsselaspekte erfolgreicher gesprochensprachlicher Interaktion
- **incrementelle/schritthaltendeVerarbeitung** ermöglicht Responsivität
  - Aktionen basierend auf partieller Information
  - flexible Abänderung geplanter und laufender Aktionen
- Verarbeitungsparadigmen und Schnittstelen müssen für inkrementelle Verarbeitung angepasst werden

Vielen Dank für Ihre Aufmerksamkeit.

Timo Baumann
baumann@informatik.uni-hamburg.de
www.timobaumann.de/work



weitere Literatur

- 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.

Raum für Notizen

Lernziele

Studierende

- ... verstehen die zwei Zeitdimensionen, die in der inkrementellen Verarbeitung betrachtet werden
- ... kennen das Konzept inkrementeller Einheiten
- ... verstehen den Vorteil, partielle und vorläufige Hypothesen systematisch zu handhaben
- ... können inkrementelle Verarbeitung auf diversen linguistischen Ebenen mit Problemen der Mensch-Computer-Interaktion in Beziehung setzen