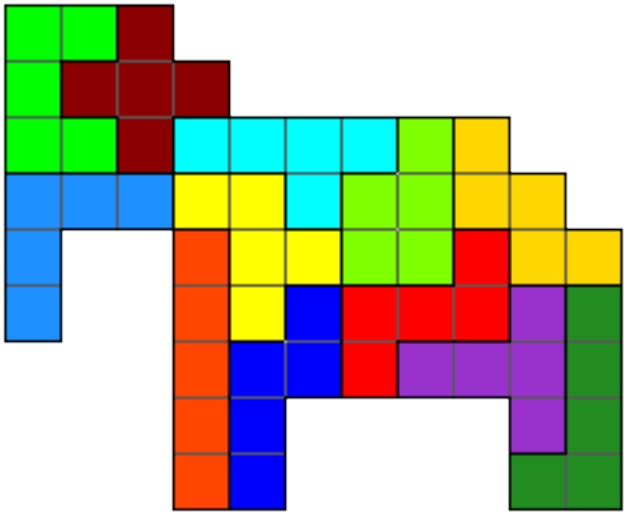
Using affordances to shape the interaction in a hybrid spoken dialog system



Timo Baumann, Maike Paetzel,
Philipp Schlesinger, Wolfgang Menzel

nats-www.informatik.uni-hamburg.de/TimoBaumann

thanks to Radu Comaneci and Mircea Pricop for helping with the implementation of the system.

Goal: A **Spoken Dialogue System**

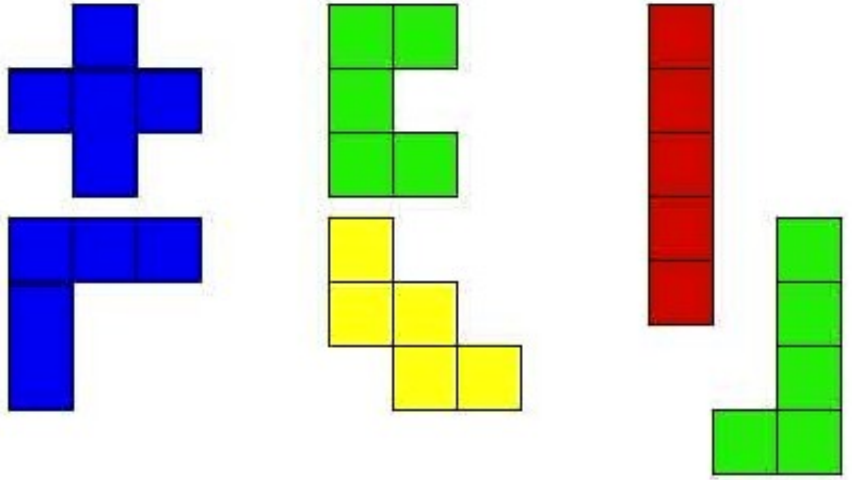
- systems that interact with humans through spoken interaction
  - task-oriented dialogue, for now
  - not necessarily as a HCI, but eventually as *companions*:

    ~~„computer, open the pod bay doors!“~~

    „hey, uh, could you help me with the .. uuh, thanks, buddy!“

Goal: A Spoken Dialogue System for a Puzzle Game Domain

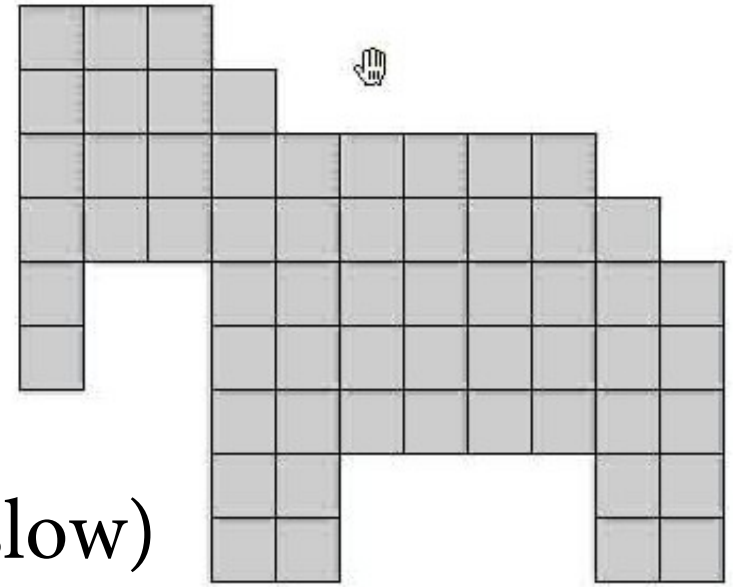
game alternates between



- selecting a puzzle piece:

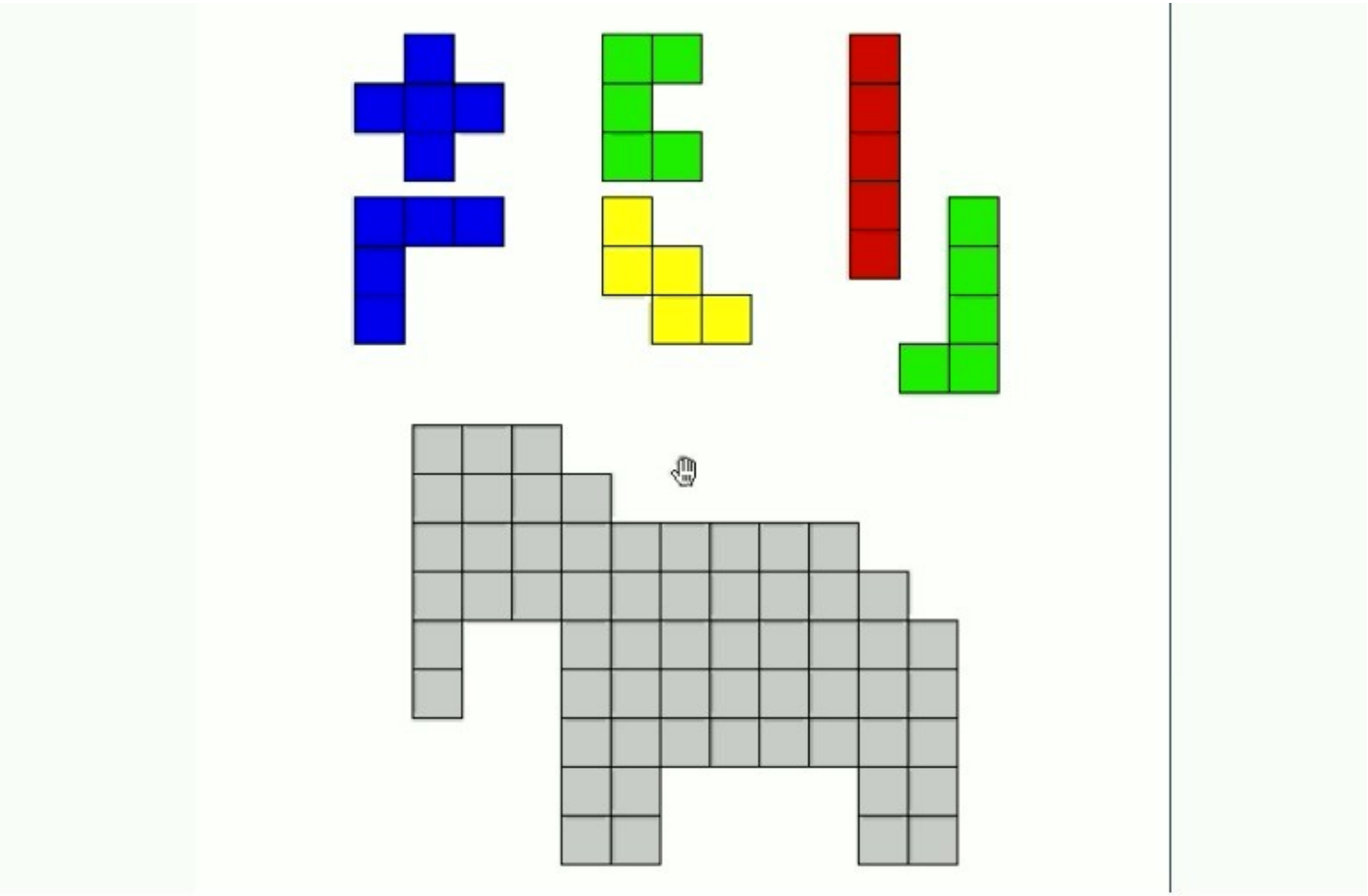
  and

- placing it in the target figure:



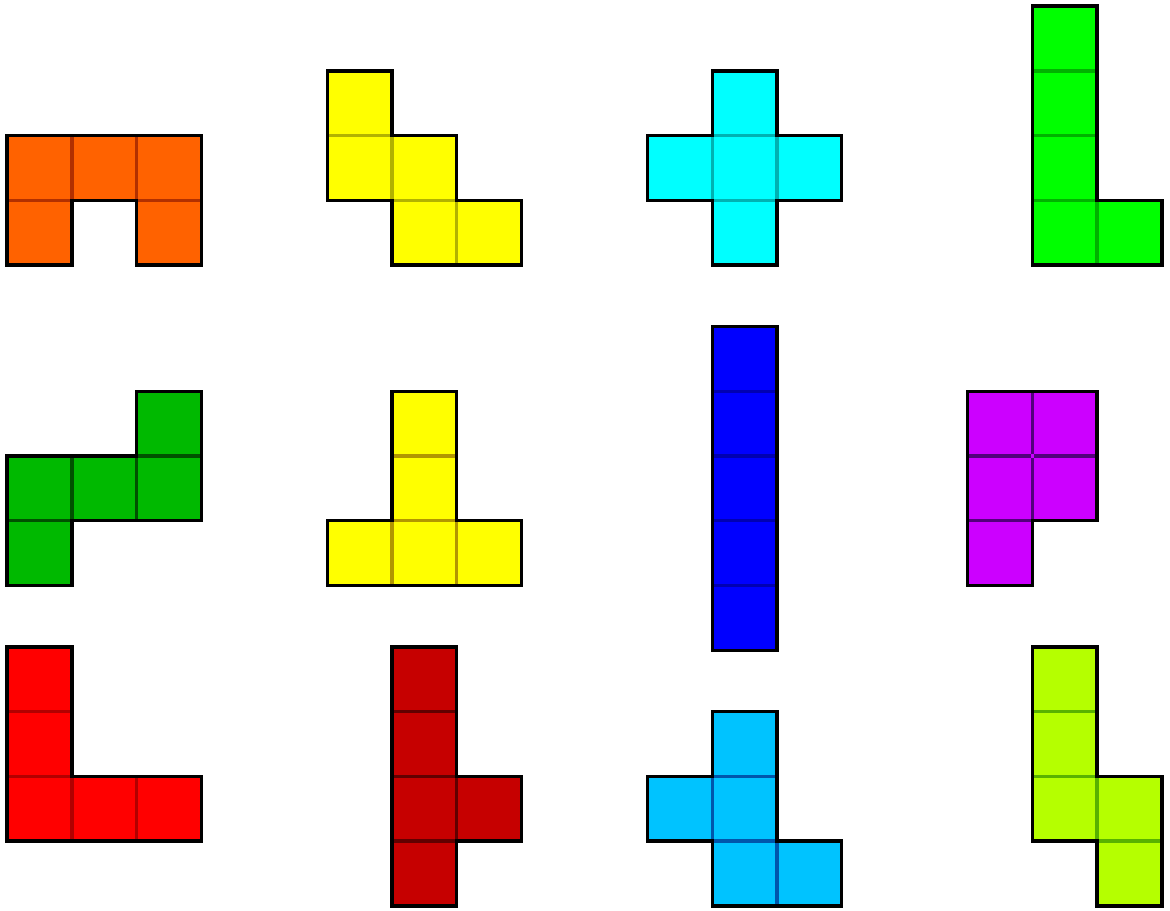
the dialogue system controls
a robot hand (that is relatively slow)

Video!

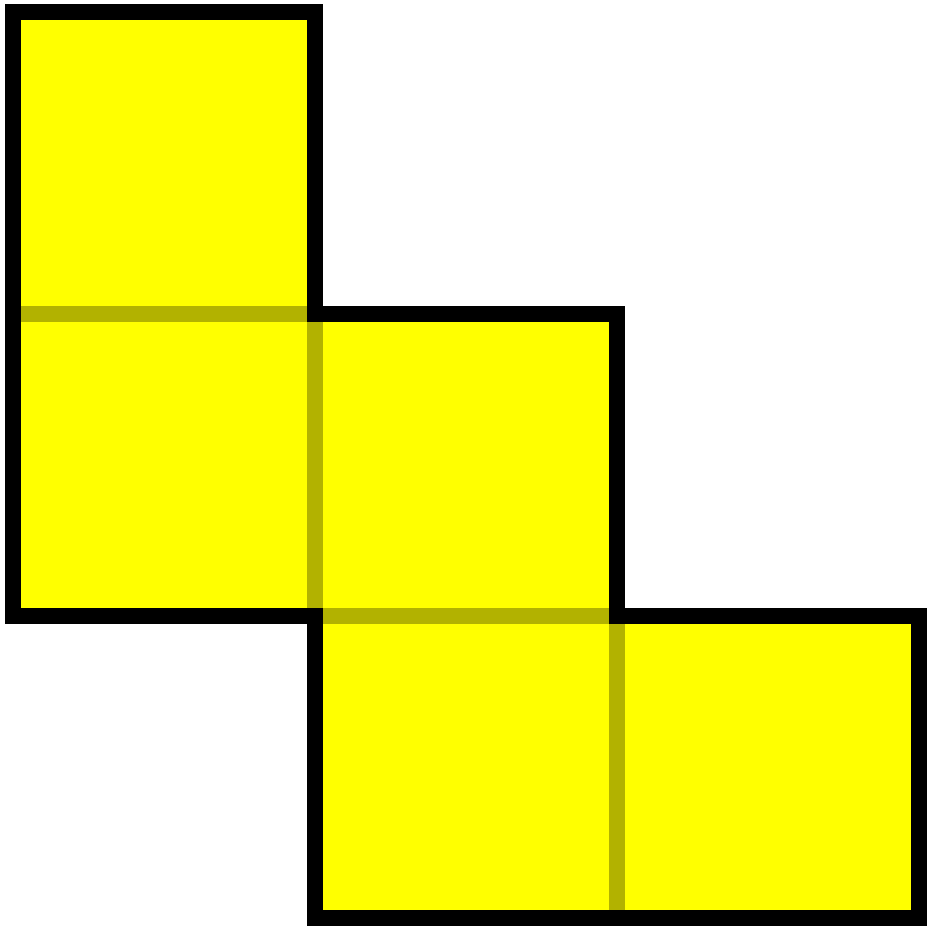


Pentomino Puzzle Pieces

all (twelve) shapes of five connected squares:



Main Challenge: Referring Expressions



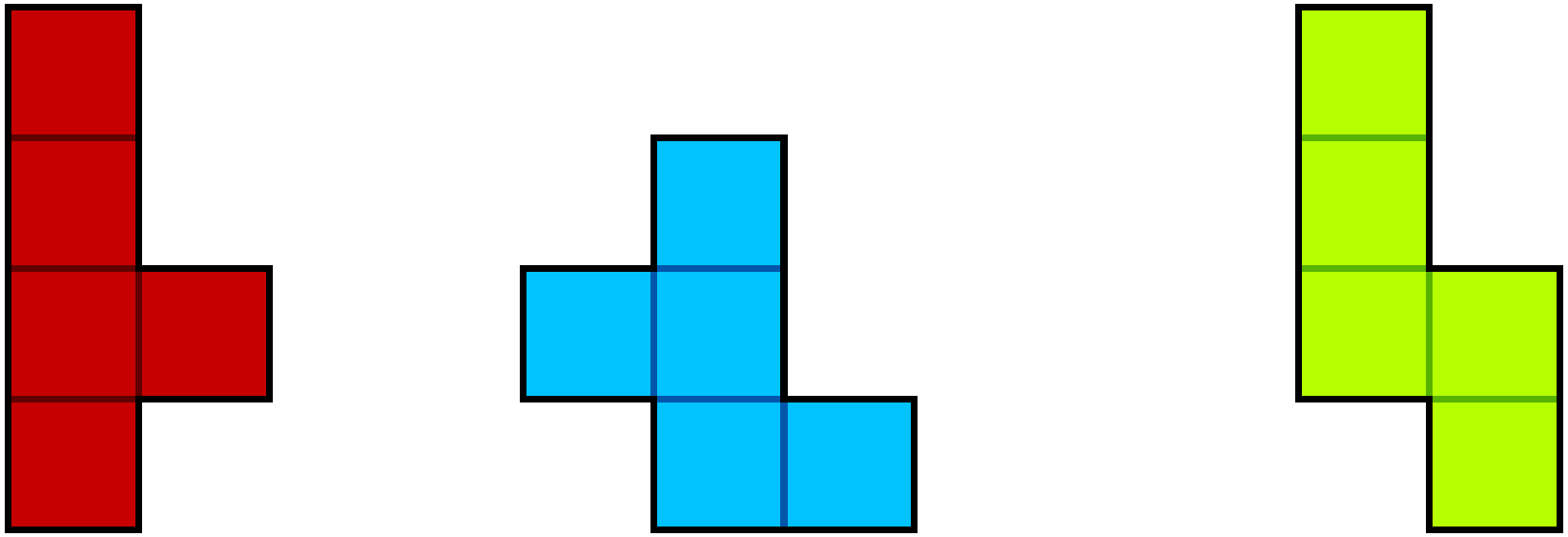
Schlangen et al. (SIGDial 2009)

Main Challenge: Referring Expressions

- large vocabulary:
  - „W, M, Treppe, Stufen, Schlange, Blitz, Croissant, Martha“
- disfluencies:
  - „das auf dem K ..... das auf'm Kopf stehende . W“
- complex expressions:
  - „besteht aus fünf Quadraten . zwei unten waagerecht dann über dem zweiten äh . noch ein Quadrat . rechts davon noch eins und darüber noch eins“
- color / relative position to other pieces, etc.

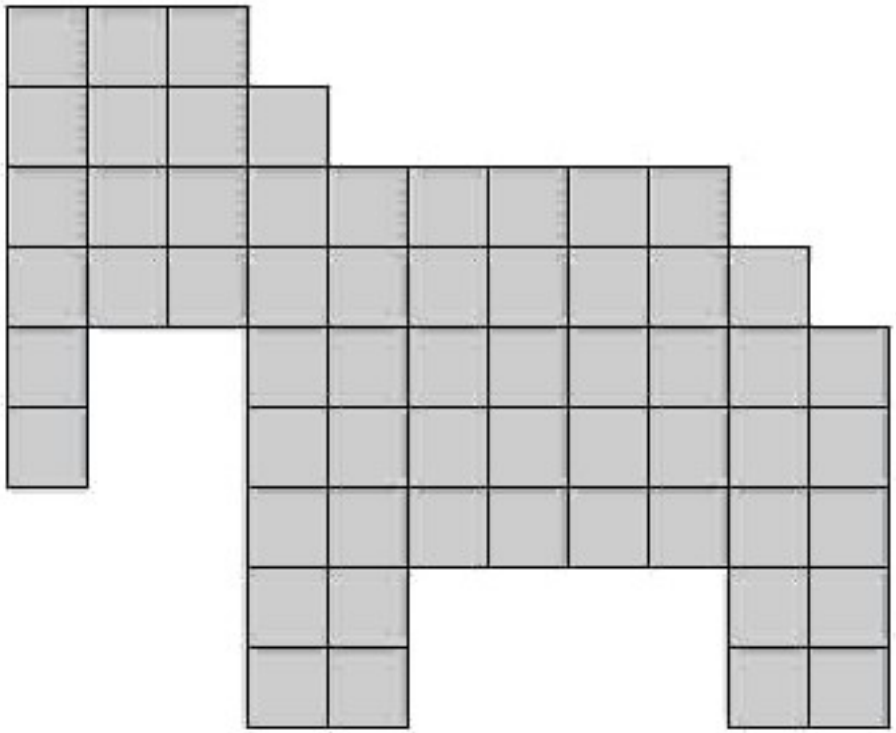
Schlangen et al. (SIGDial 2009)

that shape was easy, how about these?



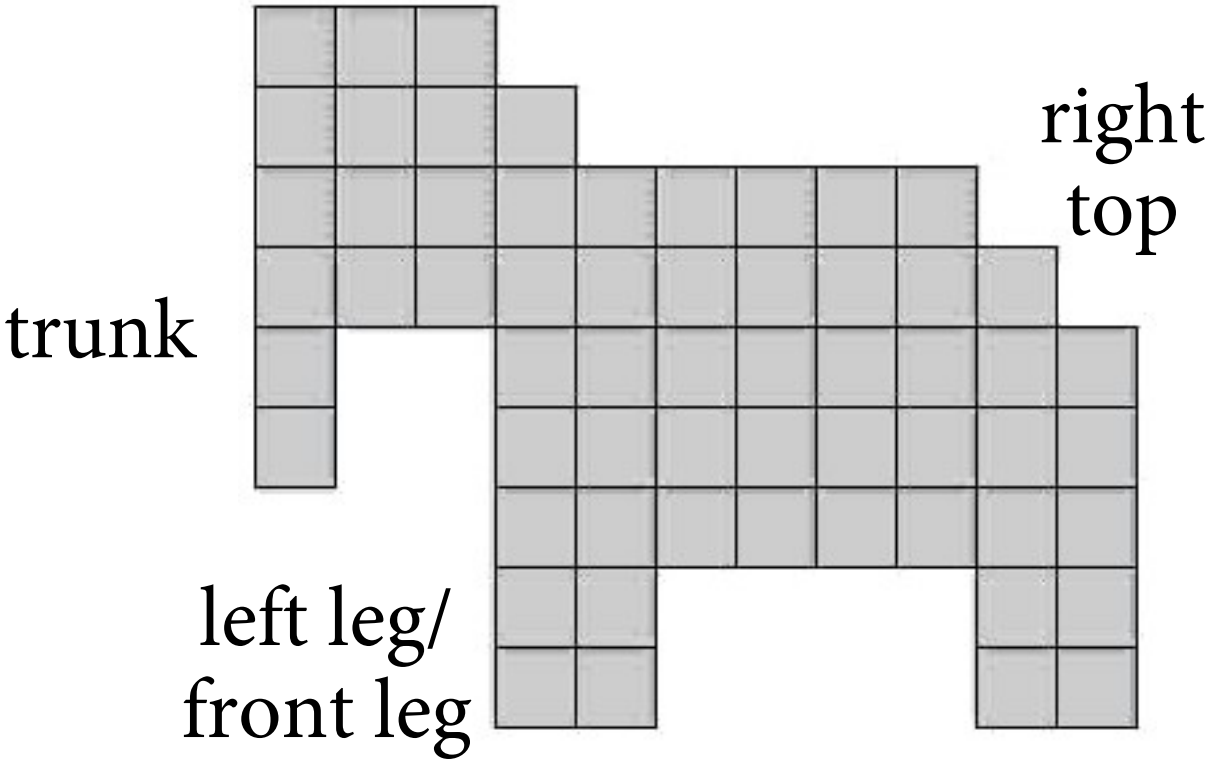
Even Worse: Referring Expressions for Targets

- just twelve pieces, but many more ways to arrange them in the figure



- even though there is just one *correct* position for every piece, all other positions can be described as well

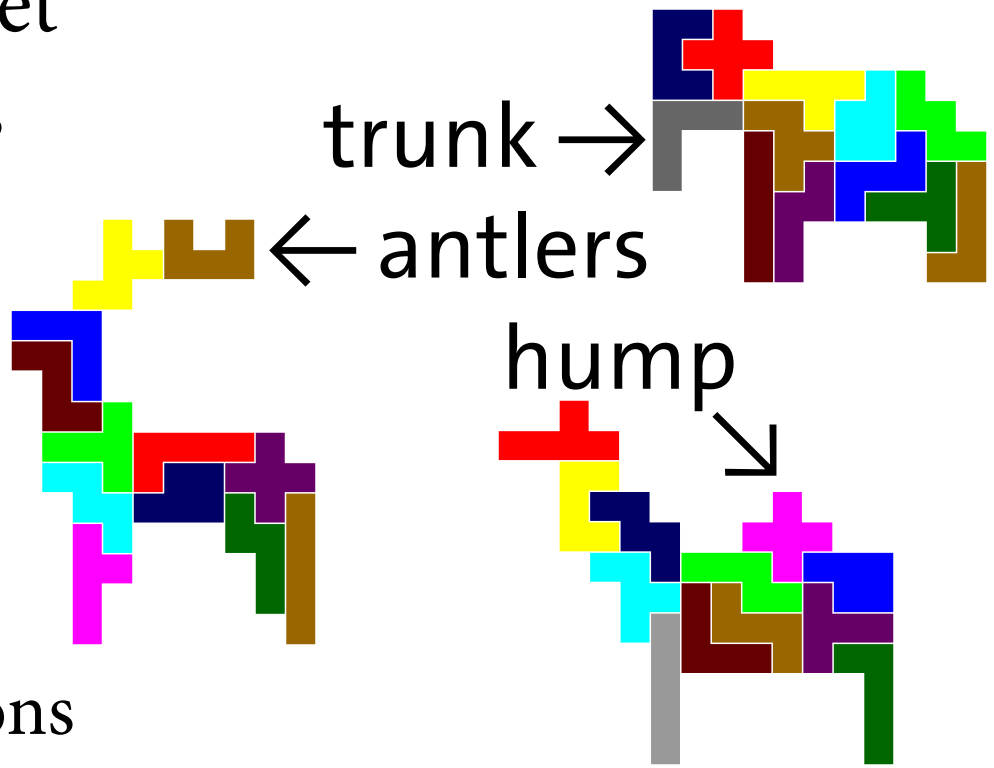
Even Worse: Referring Expressions for Targets



„also äh – ach das ist ein Elefant! – also am Rücken und da dann etwas nach rechts, dann zwei Kästchen runter und noch etwas rechts.“

Other Puzzle Figures?

there is just a fixed set of Pentomino pieces



- but an unlimited number of shapes
  - all triggering different associations for target positions

Very Hard Problem for Natural Language Processing (ASR & NLU)

- **root causes**: too little feedback, too little guidance

- our goal:
  - **move complexity** away from generating/understanding referring expressions **towards** the **interaction loop**
    - here: for puzzle piece positioning
  - **guide the user by** strategically using the **affordance of motion**

Affordances

conventionalized attribute-meaning pairs that manifest possibilities of interaction

- doors afford to be opened
- blinking cursors afford to enter text

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

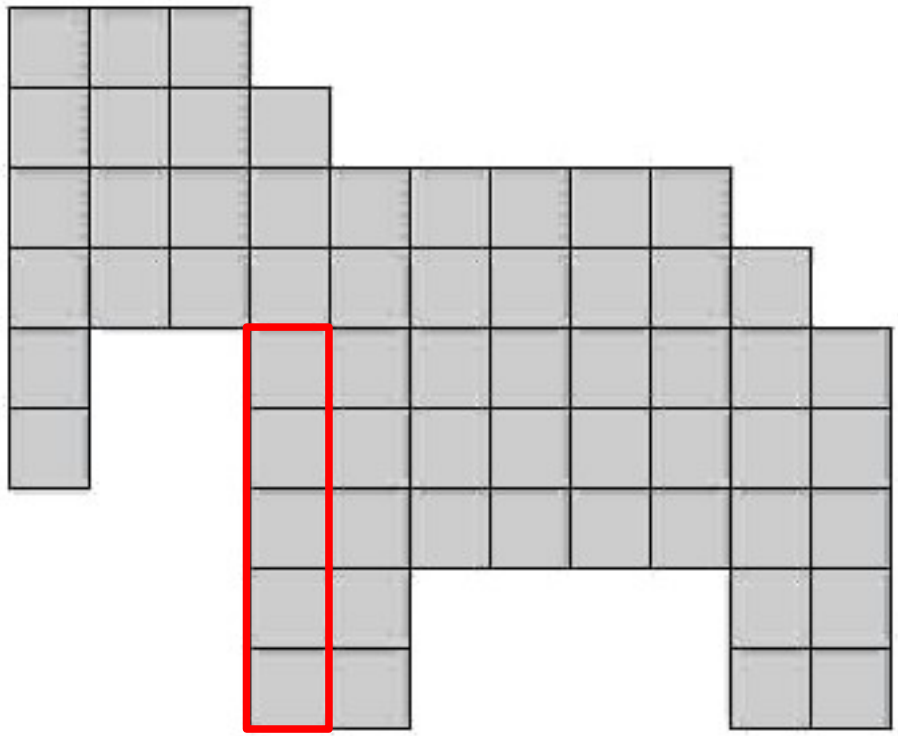
Affordances in our system

conventionalized attribute-meaning pairs that manifest possibilities of interaction

- system questions („where should I put the piece?“) afford to answer to that question
- puzzle pieces afford to be described in certain ways (by color, by shape, by position on the board, …)

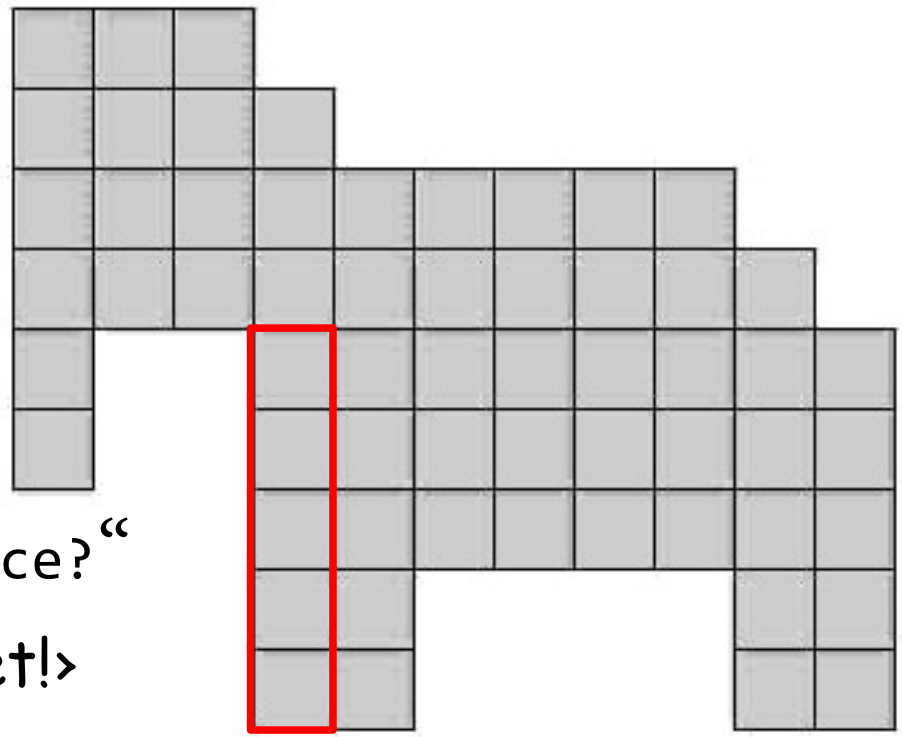
Affordances in our system (cont'd)

- figure shapes afford to have their targets described by using a certain vocabulary
- it's not the user's fault
  that they use
  weird vocabulary,
  it's the system's fault!



Affordances in our system (cont'd)

- figure shapes afford to have their targets described by using a certain vocabulary
- it's not the user's fault
  that they use
  weird vocabulary,
  it's the system's fault:



sys: „Where should I put the piece?“

usr: <I have to describe the target!>

Affordances and how they go wrong

- sys: „Where should I put the piece?“
- usr: ‹I have to describe the target!›
- usr: ‹how can I describe the target? → look at figure›
- figure: ‹well, this can be seen as a front leg. or the leg that is furthest to the left. (if you noticed that this is an elephant.) or the position in the lower left (ignoring the fact that the lower left itself is empty). or the position XY down from the top right. or …›
- usr: „Put it .. uh .. put it .. uh uh .. hm.“
- sys: ‹what's wrong with this user?›

The **Affordance of Motion**

motion manifests the possibility of interacting with the motion itself (*steering*)

- steering (in 2D) is comparatively easy:
  - four directions, go back, stop, finish.
- a system that supports steering makes its life a lot easier

- however, to keep up the steering metaphor, the system must react to commands without delay
  - otherwise the user might revert to target naming

Affordances and how they work well

- sys: „Where should I put the piece?“
- sys: ‹starts moving the piece **immediately**›
- usr: ‹I'll just control the motion!›
- usr: „Put it further to the left, go on, stop. OK.“
- sys: ‹that was easy› usr: ‹that was easy›

more Video!

Our System

- supports steering via *incremental processing* with very low delays (for positioning)
  - relatively weak speech recognizer
  - simplistic vocabulary limited to steering
- relies on standard techniques for piece selection
  - commercial grade speech recognizer
  - large grammar, numerous subdialogues for problem resolution
  - implemented with DialogOS

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)

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)

Thank you very much for your attention.

read more about this work in:

T. Baumann, M. Paetzel, P. Schlesinger, and W. Menzel: „Using Affordances to Shape the Interaction in a Hybrid Spoken Dialog System“ *Proceedings of Elektronische Sprachsignalverarbeitung* (ESSV 2013), Bielefeld, Germany, 2013.