

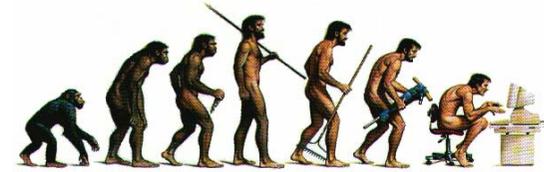
# Human-Computer Interaction

---

Termin 11:

Natural Interaction - Multimodal Interfaces

# The evolution of user interfaces

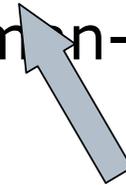


Year	Paradigm	Implementation
1950s	None	Switches, punched cards
1970s	Typewriter	Command-line interface
1980s	Desktop	Graphical UI (GUI), direct manipulation
1980s+	Spoken Natural Language	Speech recognition/synthesis, Natural language processing, dialogue systems
1990s+	Natural interaction	Perceptual, multimodal, interactive, conversational, tangible, adaptive
2000s+	Social interaction	Agent-based, anthropomorphic, social, emotional, affective, collaborative



# Natural interfaces

- Highly interactive, multimodal interfaces modeled after natural human-to-human interaction

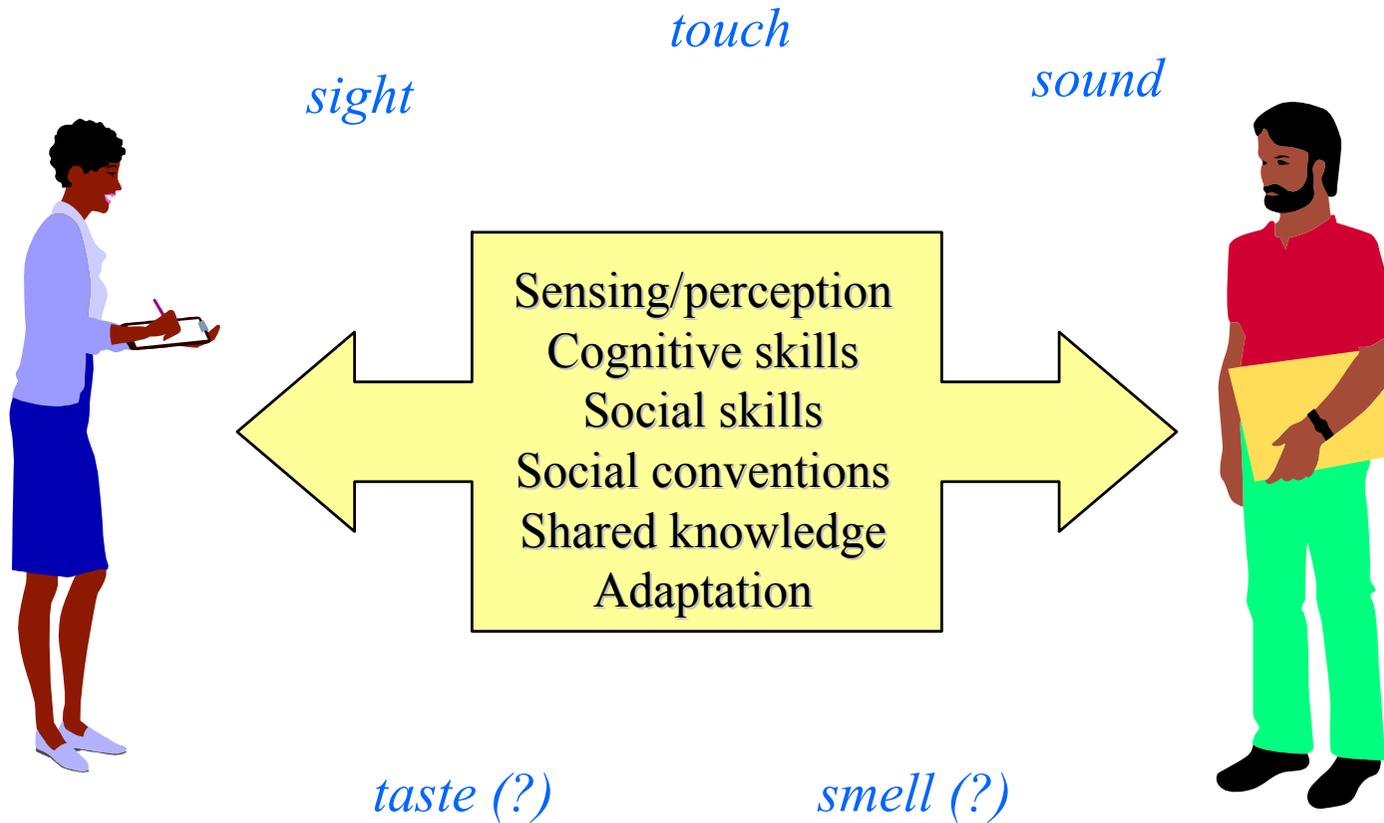


not just mouse, keyboard,  
monitor...

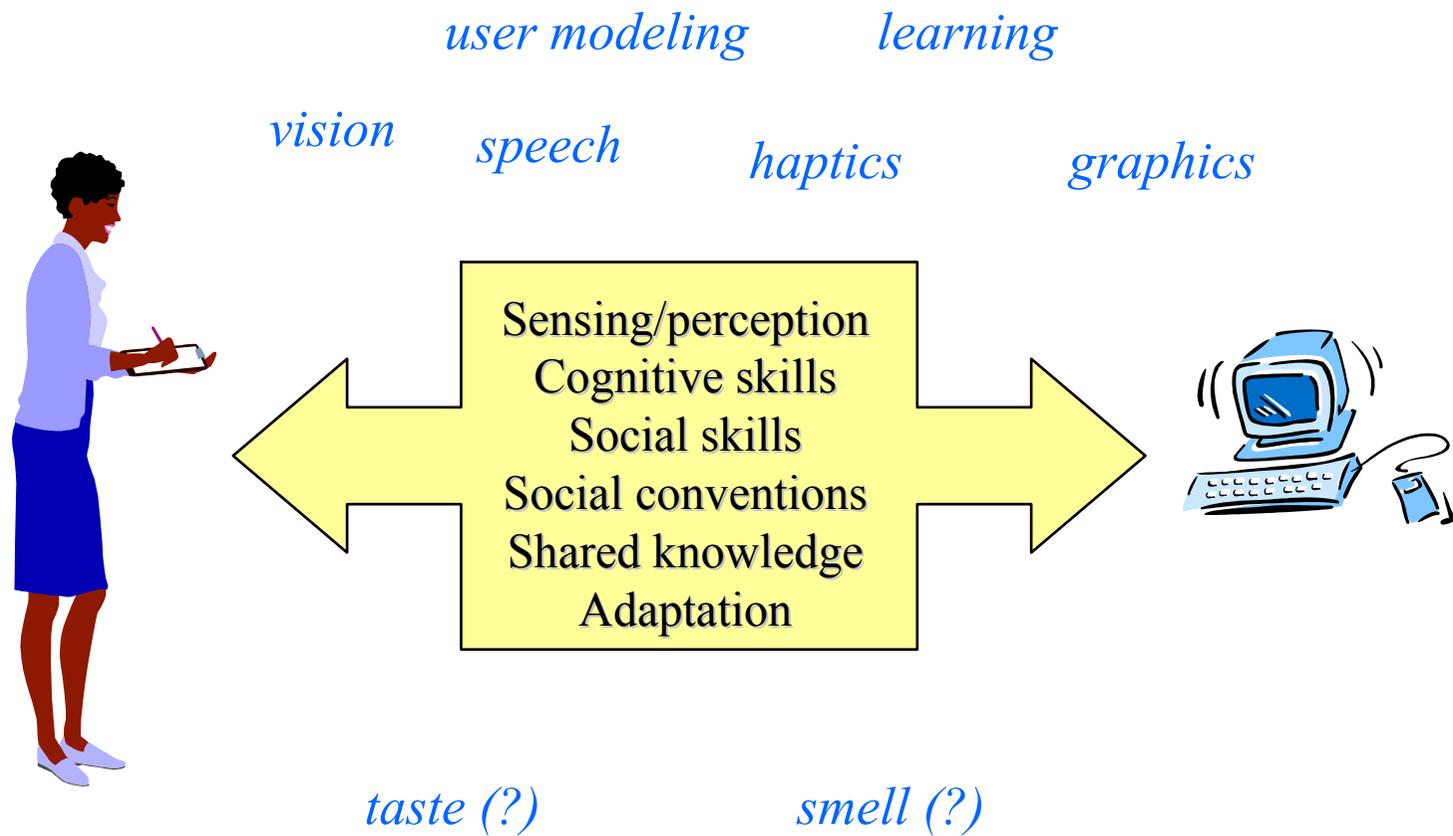
- *Goal:* For people to be able to interact with computers in a fashion similar to how they interact with each other and with the physical world



# Natural human interaction



# Natural interfaces



# Modalität

Der Begriff „Modalität“ wird unterschiedlich verwendet:

## **physiologisch**

### *sensorische Modalität*

Möglichkeiten der menschlichen Wahrnehmung:  
*visuell, auditiv, taktil, olfaktorisch, gustatorisch, vestibular*

### *motorische Modalität*

Möglichkeiten des Handelns bzw. Kommunizierens:  
*verbal, manuell, mimisch, körperlich*

## **technisch**

### *Modalität als Interaktionstechnik*

Zusammenschluß  $\langle d, L \rangle$  eines Interaktionsgeräts  $d$   
mit einer Interaktionssprache  $L$



# Modalität & multimodal

Im folgenden verwendete Definition:

*Eine **Modalität** bezeichnet ein kommunikatives System, das durch die Art und Weise wie Information kodiert und interpretiert wird, gekennzeichnet ist.*

- Modalitäten betreffen sowohl die Informationsübertragung vom Menschen zur Maschine (technisch: **Eingabemodalität**) als auch von der Maschine zum Menschen (technisch: **Ausgabemodalität**)
- Eine Schnittstelle ist **multimodal**, wenn sie mehrere Eingabemodalitäten und/oder Ausgabemodalitäten zur Verfügung stellt.



# Natürliche Modalität & Kulturtechnik

- *Natürlich* oder *fundamental* ist eine Modalität, wenn sie Teil der kommunikativen Grundkompetenz eines (sozialisierten) Menschen ist – dazu gehören: *Sprache (Laute), Gestik, Mimik, Körpersprache (Proxemik)*
- Achtung: natürliche Modalitäten sind kulturabhängig!  
Ausnahme: Emotionen, die durch Mimik, Prosodie, Körpersprache ausgedrückt werden
- Neben den natürlich vorhandenen können Modalitäten erlernt werden: Kulturtechniken wie z.B. Lesen & Schreiben oder auch *point-and-click* als "Subkulturtechnik"



# Warum multimodal?

## Natürlichkeit & Intuitivität

- stärkere Anpassung an den Menschen
- interagieren geschieht automatisch/unbewusst
- Unterschiedliche Nutzer bevorzugen verschiedene Modalitäten, bessere Akzeptanz v.a. unter ungeübten Benutzern

## Bandbreite & Effizienz der Informationscodierung

- Mehr Information pro Zeit ikommunizierbar

## Adäquatheit der Informationscodierung/Multifunktionalität

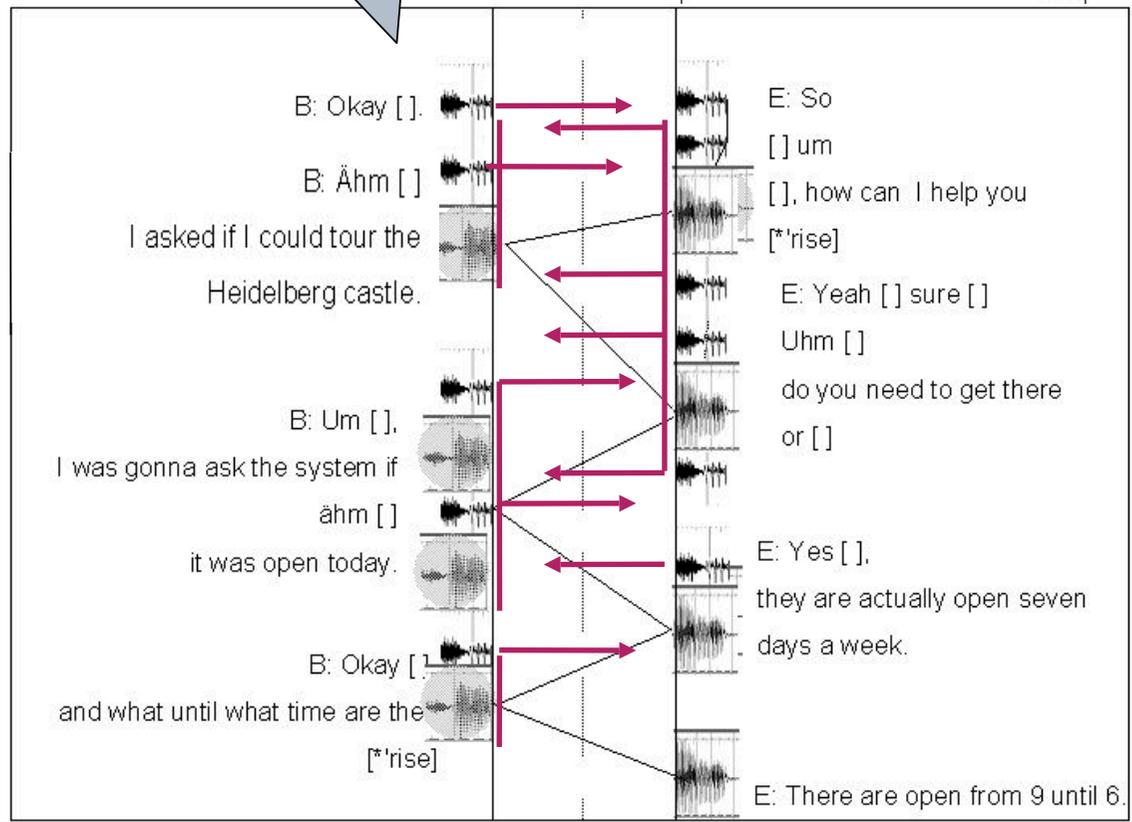
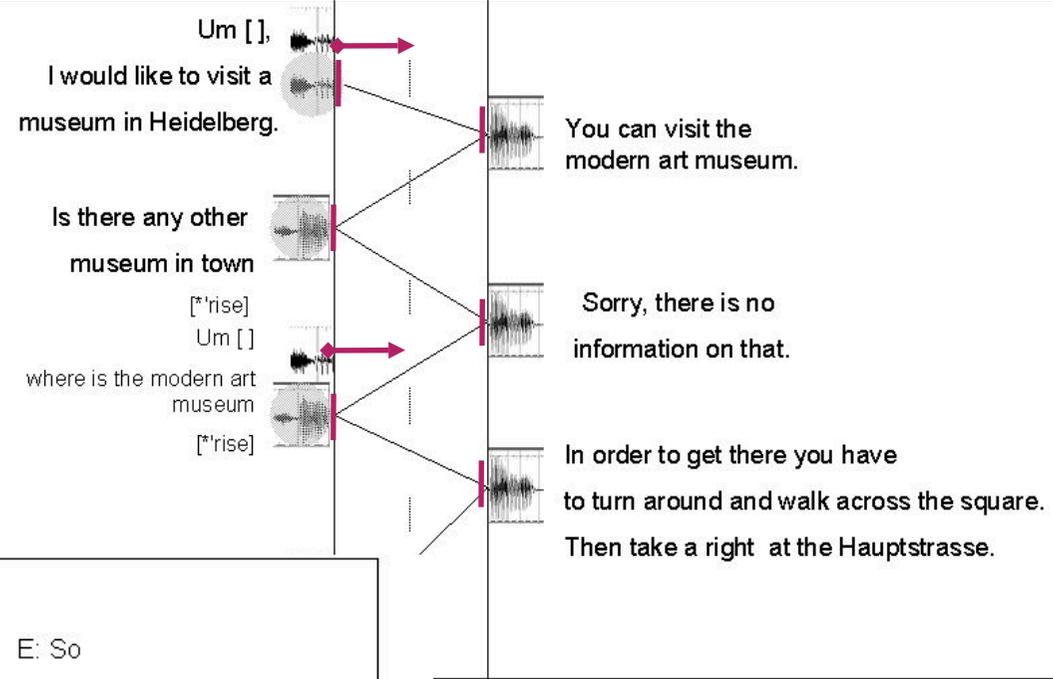
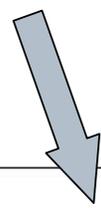
- Informationsarten/Funktionen lassen sich mit verschiedenen Modalitäten verschieden gut übermitteln/erfüllen
  - propositional (Inhalt) vs. interaktionssteuernd (Turn-taking, Feedback, Aufmerksamkeit)
  - symbolisch vs. ikonisch vs. indexikalisch

## Alternative Kommunikationswege (*universal design*)

- Berücksichtigung aller Benutzergruppen (z.B. Blinde) in allen Situationen (z.B. bei Umgebungslärm)



**In Wirklichkeit multimodal und noch komplexer (Nicken, An-/Wegschauen, Handgesten, etc.)**



### Efficient Hand-Over

- turn-taking
- turn-giving
- overlaps

# Vorteile aus Systemsicht

## Robustheit

- Weniger Überanspruchung und Abnutzung einer Modalität

## Adaptivität

- Nutzbarkeit der besten Modalität unter wechselnden Bedingungen

## Redundanz

- Übermittlung derselben Information über verschiedene Modalitäten kann die Fehlerrate verringern
- Gegenseitige Disambiguierung der Modalitäten

## Fehleranfälligkeit/-behandlung

- Nutzer wählen intuitiv weniger fehleranfälligen Modus, wechseln oft die Modalität nach Fehlerfall
- Nutzer verwenden simplere Sprache wenn sie multimodal interagieren – reduzierte Komplexität durch Aufteilung der Information

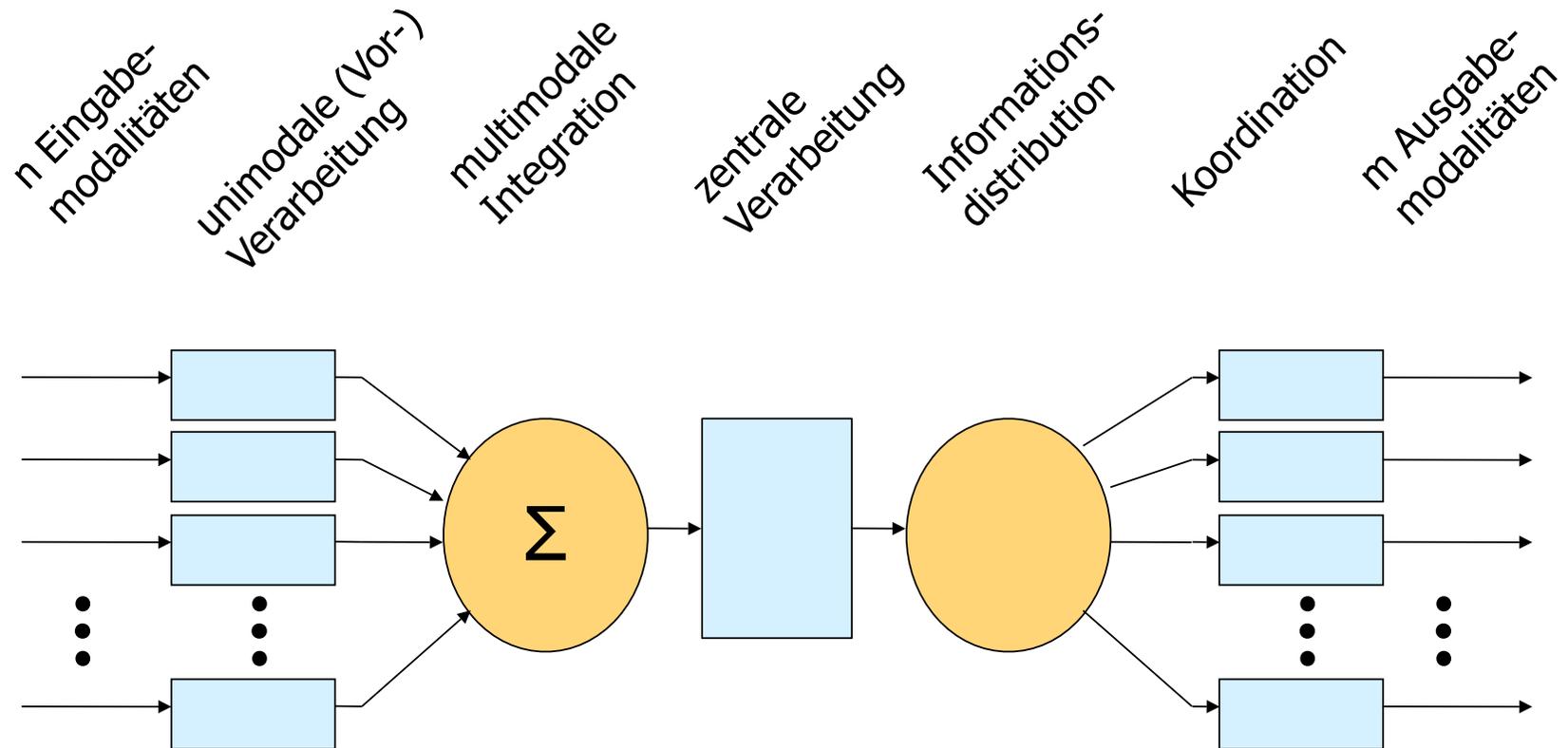


# Einwände

- HCI should be characterized by (e.g. Shneiderman):
  - Direct manipulation
  - Predictable interactions
  - Giving responsibility and sense of accomplishment to users
  
- Won't work –“AI hard”
  
- Technologische Hürde (die einfachere)
  - Lots of researchers worldwide
  - Increasing interest
  - Consistent progress
- Ökonomische Hürde (die höhere)
  - But there's growing convergence: hw/sw advances, commercial interest in biometrics, accessibility, recognition technologies, virtual reality, entertainment....



# Multimodale Systeme: Prinzip



Verarbeitungsschema für multimodale Systeme



# Multimodale Systeme: Beispiele

Das Ding ist unten rund...

Spracherkennung + Lippenlesen

Dreh das Teil so herum!

Stell den Stuhl dort hin!

Mach die Schraube da dran.



# Multimodal Interfaces vs. GUIs

## GUIs

1. Assume that there is a **single event stream** that controls event loop with **sequential** processing
2. Assume interface actions (e.g. selection of items) are **atomic** and **unambiguous**
3. Separable from application software and resides **centrally** on one machine
4. No temporal constraints, architecture not time sensitive

## Multimodal Interfaces

1. Typically process **continuous** and **simultaneous** input from **parallel** incoming streams
2. Process input modes using recognition-based technology, good at handling **uncertainty** and **ambiguity**
3. **Large computational and memory requirements**, typically distributed (e.g. multi-agent systems)
4. Require **time stamping** of input and development of **temporal constraints** on mode fusion operations



# Sprache

- *symbolische* Modalität: sprachliche Zeichen bezeichnen nur aufgrund von Konventionen
  - Ausnahme: *Onomatopoetika* (Lautmalerei)
- (gesprochene) Sprache umfaßt weitere nichtsymbolische Information: Prosodie

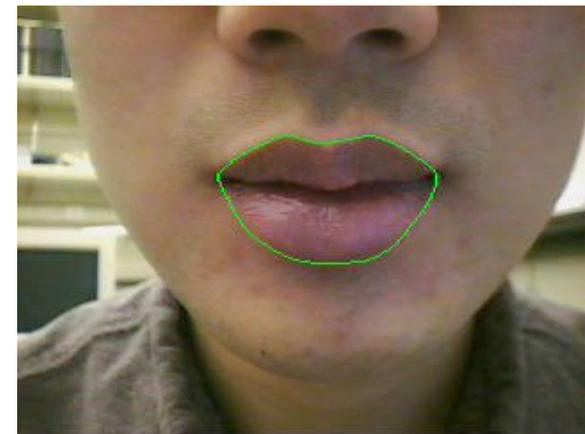
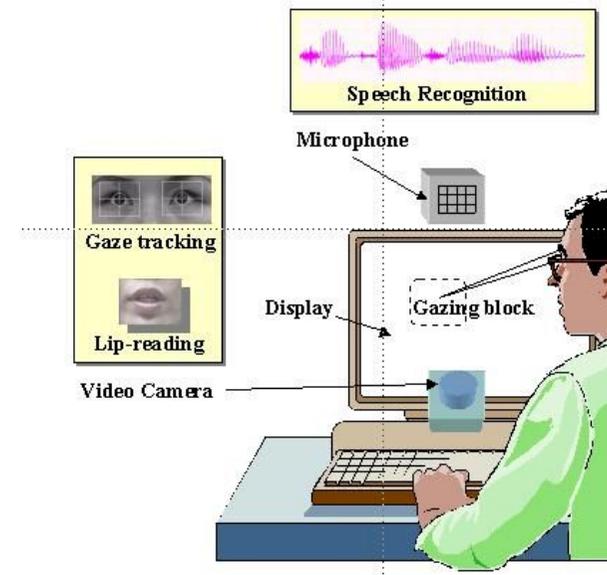
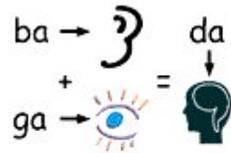
*(in der VL bereits behandelt)*



# Lippenlesen

- Sprechbewegungen des Mundes
- Ausgenutzt zur Verbesserung der Spracherkennung, vor allem bei Hintergrundgeräuschen (z.B. im Auto)

■ Erinnerung: „McGurk-Effekt“



Bimodal speech rec.,  
Rockwell Scientific Comp.

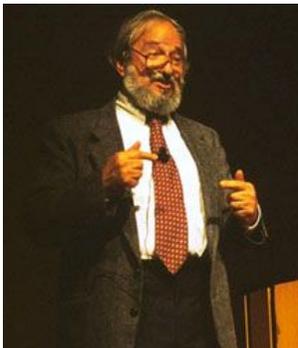


# Gestik

## □ *Kommunikative Gestik*

- Nicht manipulativ (z.B. Haare aus Gesicht streichen)
- Unmittelbar bedeutsam (z.B. nicht nervöses Wippen)

Gesten sind Bewegungen der oberen Gliedmaße,  
die aufgrund einer Kommunikationsabsicht entstanden sind.



**ikonische Geste**  
äußere Form der Geste  
ähnelt dem Objekt



**indexikalische (deiktische)  
Geste**  
verweist auf ihr Objekt im  
Kontext

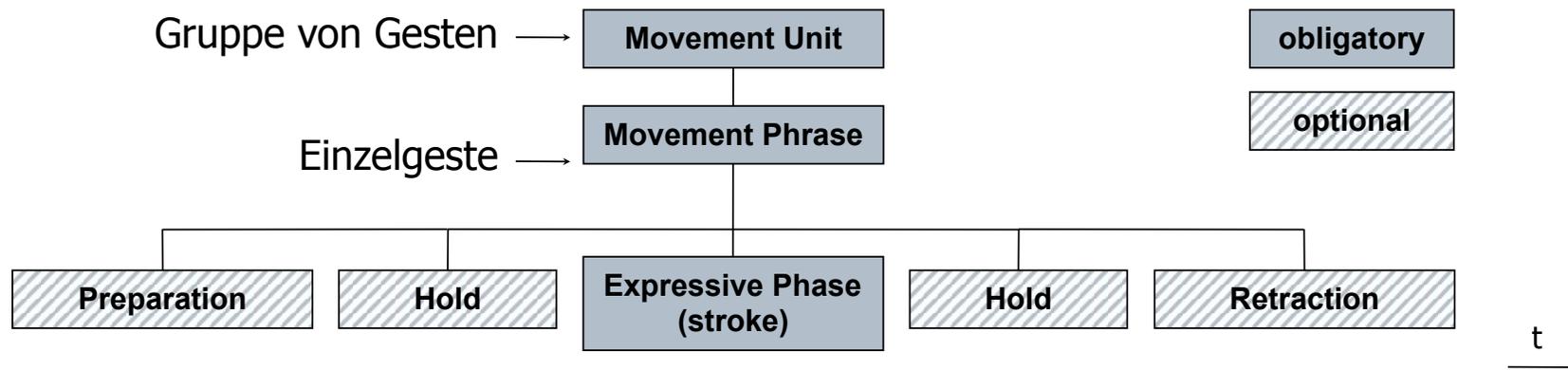


**symbolische  
(emblematische) Geste**  
konventionalisiert  
innerhalb einer  
Gemeinschaft



# Gesten: Struktur

Gesten bestehen typischerweise aus mehreren Bewegungsphasen



- Vorbereitungsphase (preparation): Hände in Stellung bringen
- expressive Phase (stroke): bedeutungstragender Teil
- Rückführungsphase (retraction): Hände zurück in eine Ruhestellung
- Haltephasen (hold): keine Bewegung



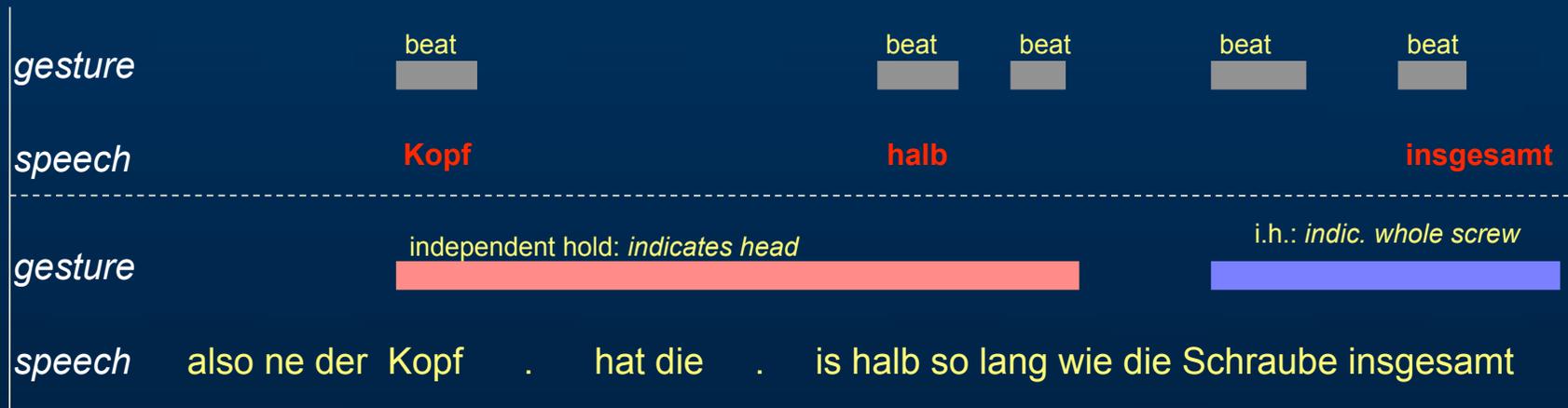


U. Twente, Dec. 2005

## Other functions of gesture

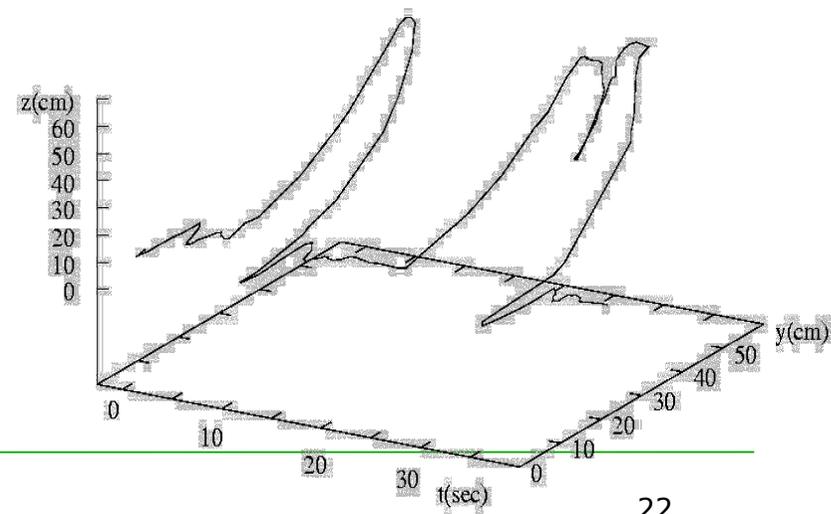
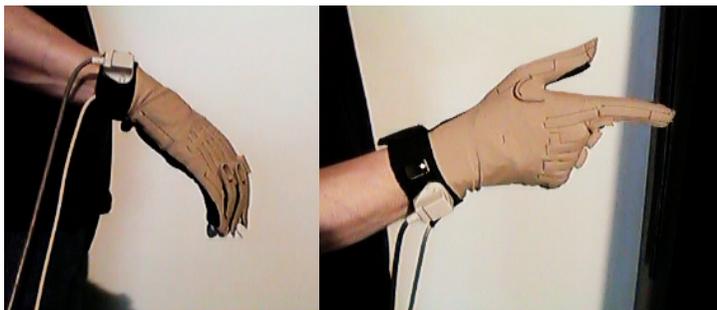
### Reflect discourse structure

- Convey, and thus mark, discursively focal elements
- Emphasize
  - Beats or beat-like movement qualities



# Gestenerkennung

- Techniken: kamerabasiert, aktives Tracking (Datenhandschuhe, Sensoren) oder passives Tracking (markerbasiert) (siehe VL zu „Input Devices“)
- *Segmentierungsproblem*:  
Wie können aus dem Bewegungsfluß die *Strokes* segmentiert werden?
- Möglichkeiten: Ausnutzung von Merkmalen wie Handspannung, Symmetrien, Stopps, etc.



# Natürliche Multimodalität: Gestik und Sprache

Zwischen Sprache und Gestik besteht ein zeitlicher und inhaltlicher Zusammenhang – zusammengefaßt in drei "Regeln" (McNeill, 1992):

- **Phonologischer Synchronismus**  
Der *Stroke* einer Geste eilt der betonten Silbe voraus oder ist mit ihr synchronisiert.
- **Semantischer Synchronismus**  
Sprache und Gestik stellen dieselbe Bedeutung zur selben Zeit dar.
- **Pragmatischer Synchronismus**  
Wenn Sprache und Gestik gemeinsam auftreten, dann erfüllen sie dieselbe pragmatische Funktion.

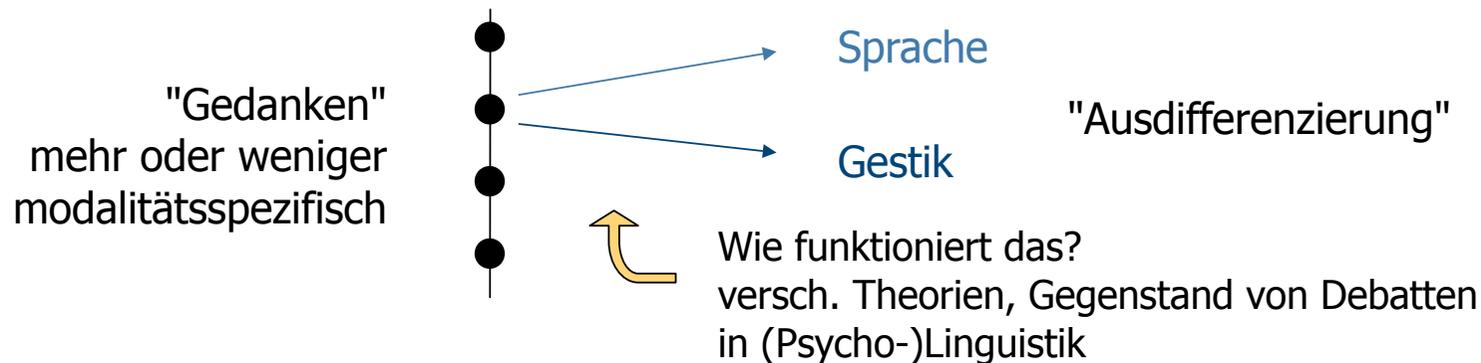


# Gestik und Sprache

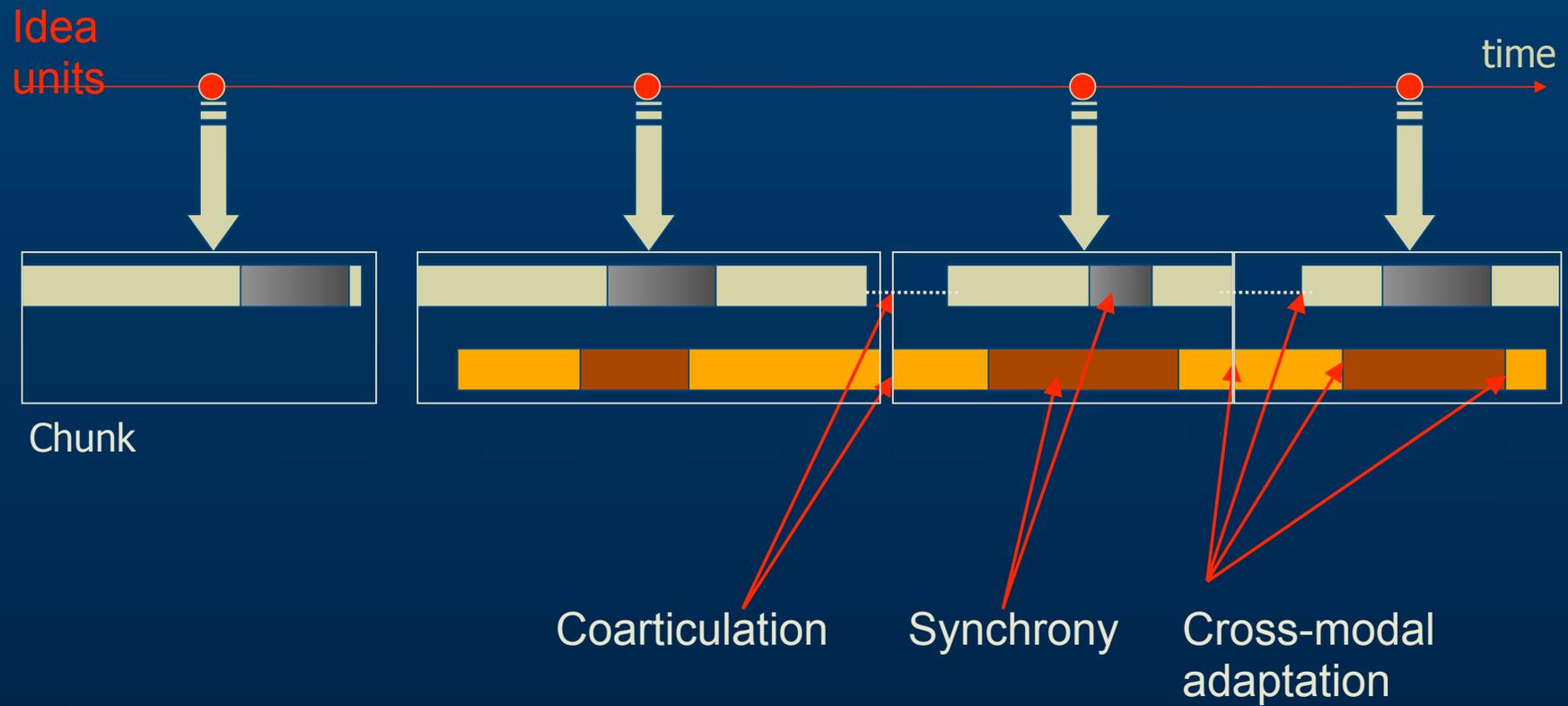
Die starke Verzahnung von Sprache und Gestik führte zu der Theorie, daß koverbale Gestik und Sprache einer gemeinsamen kommunikativen Grundidee entspringen.

## Vorstellung

Kommunikation: Abfolge von zu kommunizierenden Inhalten oder Gedanken. Gedanken entwickeln sich dann zu (oder werden verpackt in) Sprache und Gestik.



# Overall production of speech and gesture



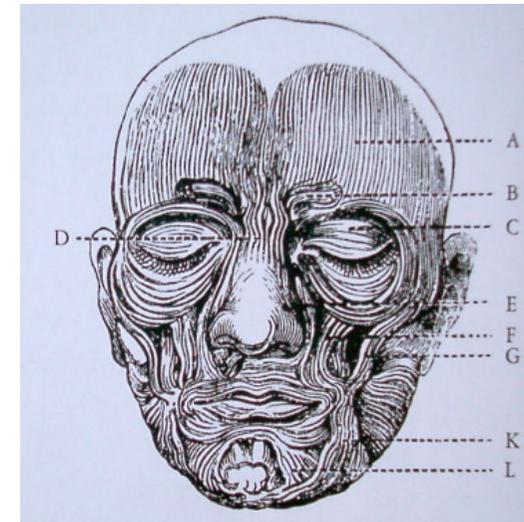
# Mimik

## Lexikondefinition (Duden)

*"Gebärden- und Mienenspiel [des Schauspielers] als Nachahmung fremden oder als Ausdruck eigenen seelischen Erlebens"*

## Biologisch

- Verschiebung der Gesichtshaut durch das Zusammenspiel verschiedener Muskeln
- auch bei anderen Primaten ausgeprägt, aber Menschen haben leistungsfähigste Mimik (feinere Muskeln als z.B. Schimpansen)



# Mimik: Emotionaler Ausdruck

Mimik trägt zum Ausdruck *emotionaler Zustände* und kommunikativen Feedbacks bei

## □ Darwin:

- Kleinkinder: Wut, Angst, Zuneigung, Freude, Neid, Schüchternheit, Unbehagen
- + bei älteren Kindern "kognitivere" Emotionen: Scham, Trauer, Verlegenheit, Resignation

## □ Heute in der Regel 6 „Grundemotionen“

- Freude, Trauer, Ekel, Überraschung, Wut, Angst
- Die Mimik dazu ist bei allen Menschen gleich

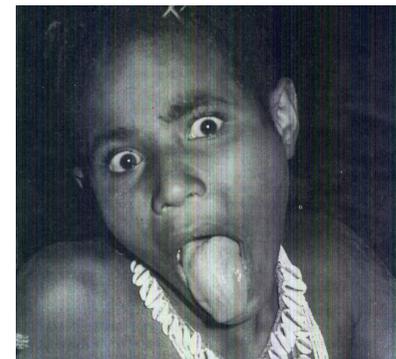
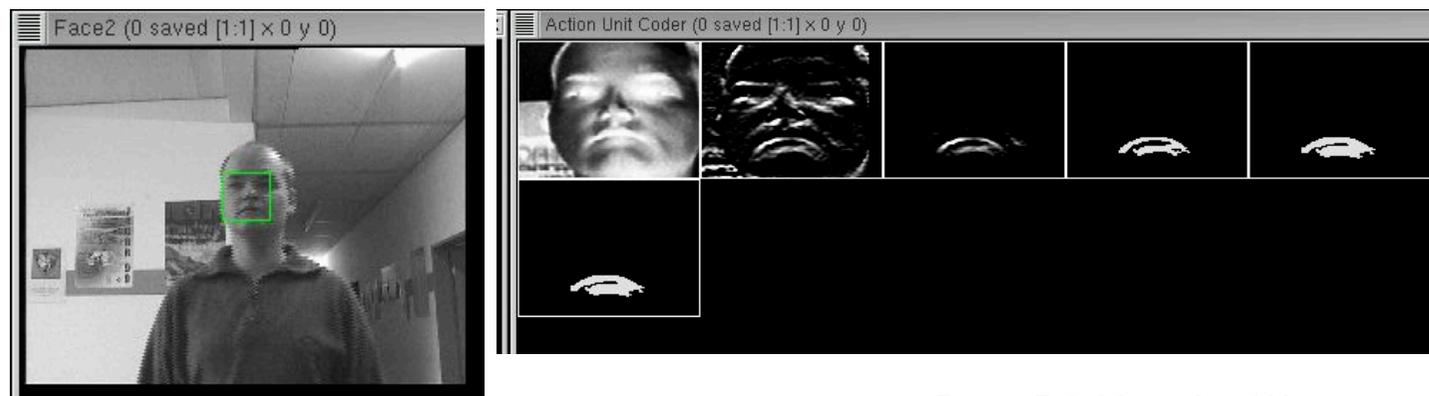


Abb.: Schiefenhövel et al.,  
1994



# Mimikerkennung

- Gesichtserkennung im Videobild
- Merkmalsextraktion: Finden bestimmter Strukturen (Augenbrauen, Augen, Nase, Mund) und Bestimmung markanter Punkte darauf
- Klassifikation der Gesichtsmimik: häufig auf Basis der Bewegungen der markanten Punkte relativ zueinander
  - Klassifikation in Emotionen (freudig, ärgerlich, ...) oder
  - Klassifikation in aktivierte Aktionseinheiten

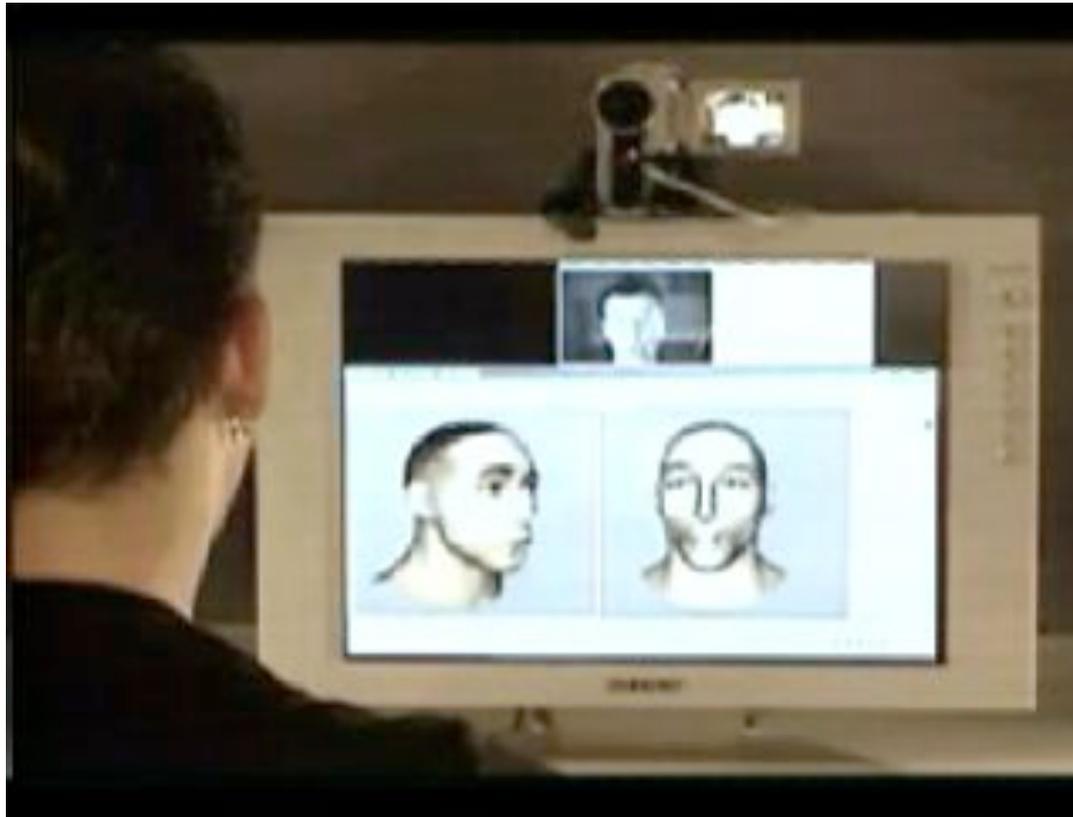


Bsp.: DA Matthias Weber



# Mimikerkennung

## □ *Facial Feature Finding & Tracking*



[www.nevenvision.com](http://www.nevenvision.com)



# Blickrichtung

- kann als "eigene" Modalität angesehen werden
- wichtig zur Aufmerksamkeitsfokus-Bestimmung oder zur Dialogsteuerung (*Turn-Taking*)
- kann interne Zustände widerspiegeln:
  - Augen beim Gespräch nach oben: nachdenken oder Gedächtnisinhalte abrufen
  - Augen nach oben + leicht geöffneter Mund: "was für ein Idiot..."

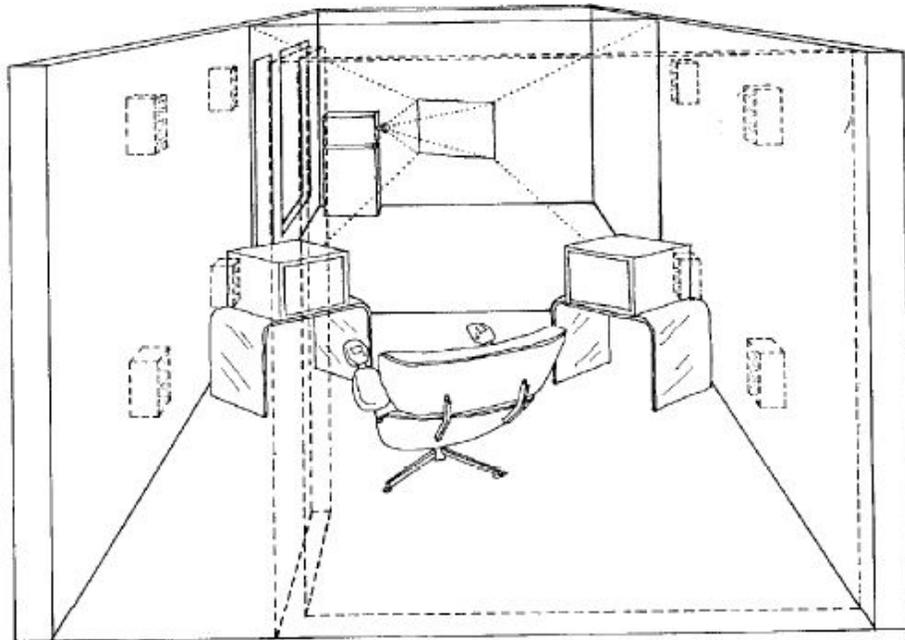


# Multimodal input processing

- The **processing** and **integration** of multiple input modalities for the communication between a user and the computer.
  
- Examples:
  - Speech and pointing gestures (Put-That-There, CUBRICON, XTRA, etc.)
  - Eye Movement based Interaction (Jacob, 1990)
  - Speech, gaze and hand gestures (ICONIC, Virtuelle Werkstatt)
  - Speech and Lip Movement



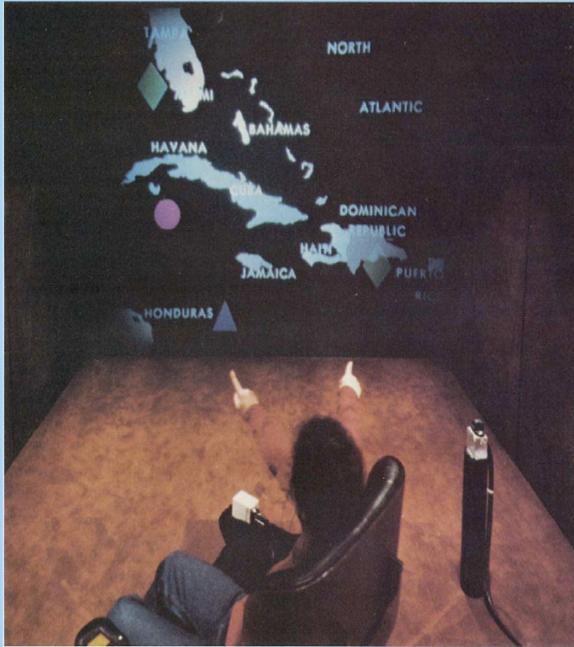
# MIT Media Room



- loudspeakers, frosted glass projection screen, TV monitors on either side of user's chair
- chair arms with one-inch high joystick sensitive to pressure and direction, touch sensitive pad
- Position-sensing cube attached to wristband



# Put-That-There (Bolt, 1980)



(Graphic taken from [1])

## **“Create”:**

“Create a blue square there.”

## **“Move”:**

“Move the blue triangle to the right of the green square”

“Move that there”

*(User does not even have to know what “that” is.)*

## **“Make that ...”:**

“Make that blue triangle smaller”

“Make that smaller”

“Make that like that”

## **“Delete”:**

“Delete that green circle”

“Delete that”



# Processing of commands

## **"Create a blue square there."**

- Effect of *complete* utterance is a "call" to the *create* routine that needs the object to be created (with attributes) as well as x,y position input from wrist-borne space sensor.

## **"Call that ...the calendar"**

- Recognizer sends code to host system indicating a naming command ("call") → x,y coordinates of item signal are noted by host → host switches speech recognition to training mode to learn the (possibly new) name to be given to the object

All utterances processed with hard-wired procedural semantics



# Multimodal analysis (revisited)

Two central problems to be solved (Srihari, 1995):

## segmentation problem

*how can a system be made to cope with 'open input'?  
how can continuous input be segmented into units that  
can be processed in one system cycle?*

## correspondence problem

*how to determine what relates to what across the  
multiple input modalities?*



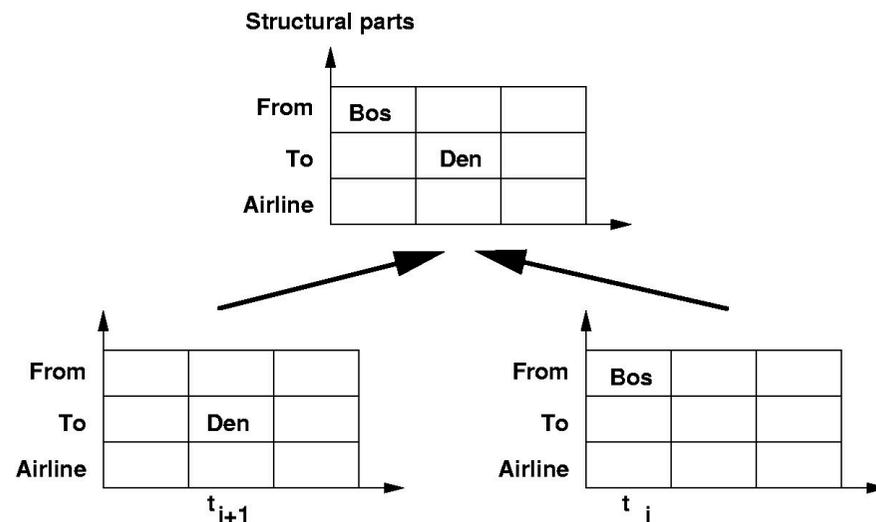
# Multimodale Integration (Fusion)

- Zu beachten:
  - **zeitliche** Zusammenhänge  
Bsp.: *"stell dieses <Zeigegeste> Ding dort hin"*  
→ Bezieht sich die Geste auf das Objekt (dieses) oder den Ort (dort)?
  - **Semantisch-pragmatische** Zusammenhänge  
Bsp: „drehe diese <ikonische Geste> Leiste so herum“  
→ Geste deutet Drehung an, bezieht sie sich auf Objekt oder Aktion?
  
- Typische Methode: Übernahme und Erweiterung von Techniken aus dem Bereich des Parsens natürlicher Sprache (*"Multimodale(s) Grammatik/Parsen"*)



# Integration mit Framestrukturen

- Modellierung einer Benutzerinteraktion als eine *Frame* mit festen *Slots* von Attribut-Wert-Paaren
- Modalitäten können einen oder mehrere *Slots* füllen bis Struktur komplett spezifiziert ist
- Nachteil: feste Struktur; erlaubt nur einen Typus von Interaktion



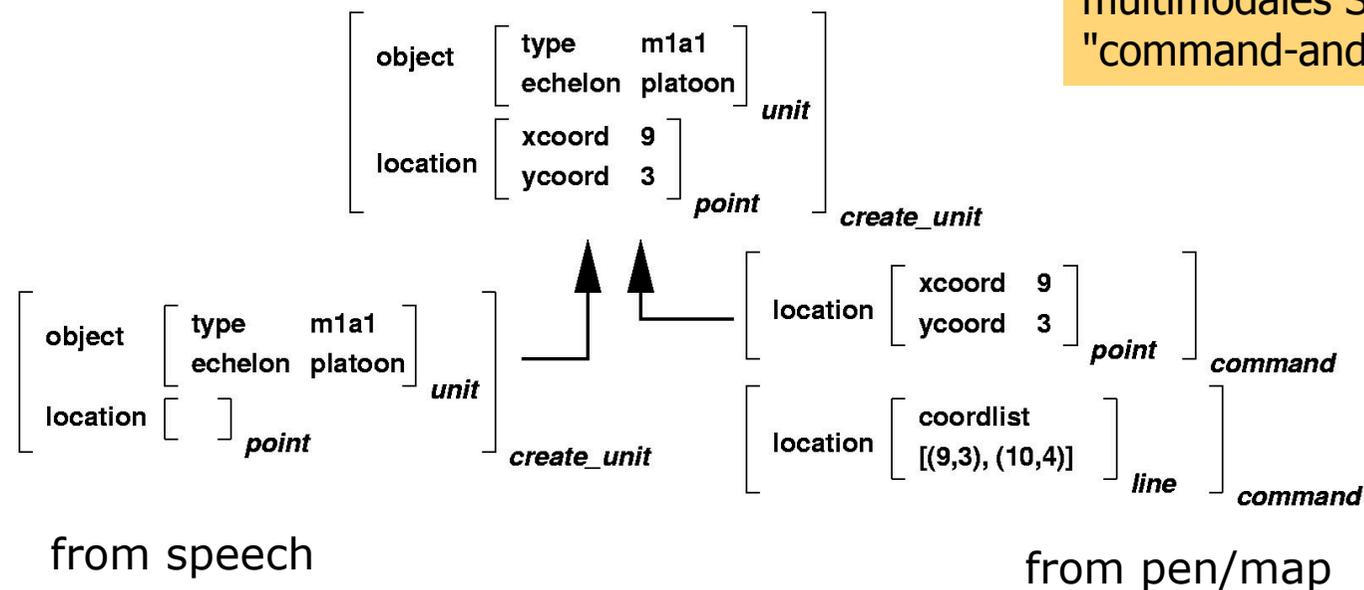
Beispiel: **MATIS**,  
**M**ultimodal **A**ir **T**ravel  
**I**nformation **S**ystem

(„Melting pots“)

# Integration mit getypten Attribut-Wert-Strukturen

- im Gegensatz zu starren Frame-Strukturen Einführung unterschiedlicher Frame-Typen
- Integration durch Unifikation der Strukturen

Beispiel: **QuickSet**,  
multimodales System für  
"command-and-control"

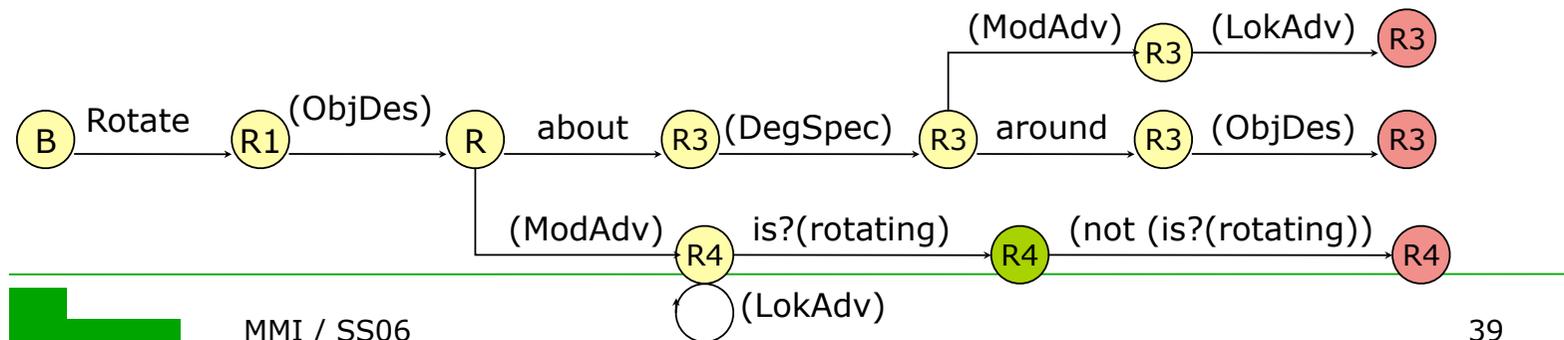


# Integration mit Automaten/ Übergangsnetzen

- Parsen eines multimodalen Ausdrucks durch Zustandsübergangsnetze (STN, ATN)
- Alphabet der Eingabesymbole z.B.: Menge von Worten und Menge verschiedener Gesten
- Problem: Im Gegensatz zu Sprache sind multimodale Eingaben (etwa Sprache und Gestik) nicht sequentiell; Möglichkeiten zur Flexibilisierung der starren Abfolge von Eingabezeichen nötig

Beispiel: **tATN**

„Rotate [pointing] this thing about 30 degrees to the right.“  
„Rotate the yellow wheel like [rotating] this.“



MMI / SS06

39

(Latoschik, 2001)

# CUBRICON (Neal & Shapiro, 1991)

- System integrating deictic and graphic gestures with simultaneous NL for *both* user input and system output
- interface capabilities
  - Accepts and understands references to entities in NL & pointing
  - Disambiguates unclear references and infers intended referent
  - Dynamically composes and generates synchronous spoken NL, gestures and graphical expressions in output

Calspan-UB Research  
Center Intelligent  
CONversationalist

MMI / SS06

## Cubricon Dialogue Example

■ user: *what aircraft are appropriate for the mission?*

■ system:

What AC are Approp...	TEXT	PRINT
F4C	F4D	F4E
F4G	F111E	F111F
F-15	F-15E	F-16

■ user: *<click on F4C in table> what is its speed?*

■ system: *An F4C has a speed of 260 metres per second. No F4Cs are stationed at Alconbury*

■ user: *What are the speeds of the planes?*

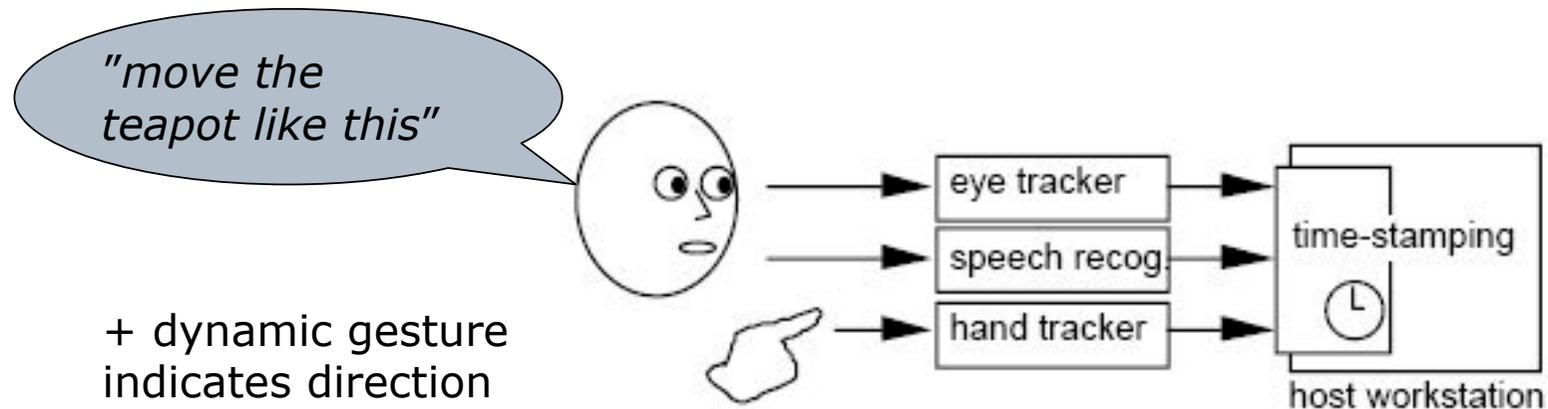
# CUBRICON Knowledge Sources

- Multimedia parser: ATN network for NL + mouse gesture
- Used in understanding input and generating output
- Knowledge Sources:
  - **Lexicon**
  - **Grammar**: defines multimodal language
  - **Discourse Model**: Representation of “attention focus space” of dialogue. Has a **focus list** and **display model** – tries to retain knowledge pertinent to the dialogue
  - **User Model**: Has dynamic “Entity Rating Module” to evaluate relative importance of entities to user dialogue and task – tailors output and responses to user’s plans, goals and ideas
  - **Knowledge Base**: Information about task domain, all objects and concepts represented in a single knowledge representation language (semantic net-based)



# ICONIC (Koons et al., 1993)

- Integrating simultaneous speech, gestural, and eye movement (for reference resolution for map and blocks world interaction)
- Problems: timing and abstraction
  - All three streams of data are collected on a central workstation and assigned time stamps, used later to realign data



# ICONIC: Processing input streams

## Step 1 - Parsing

- Parse input data stream
- Generate frame-based description of the data

## Step 2 - Evaluation

- Encode and evaluate the frames based on two models
- Every frame has method that controls search for frame values in KB

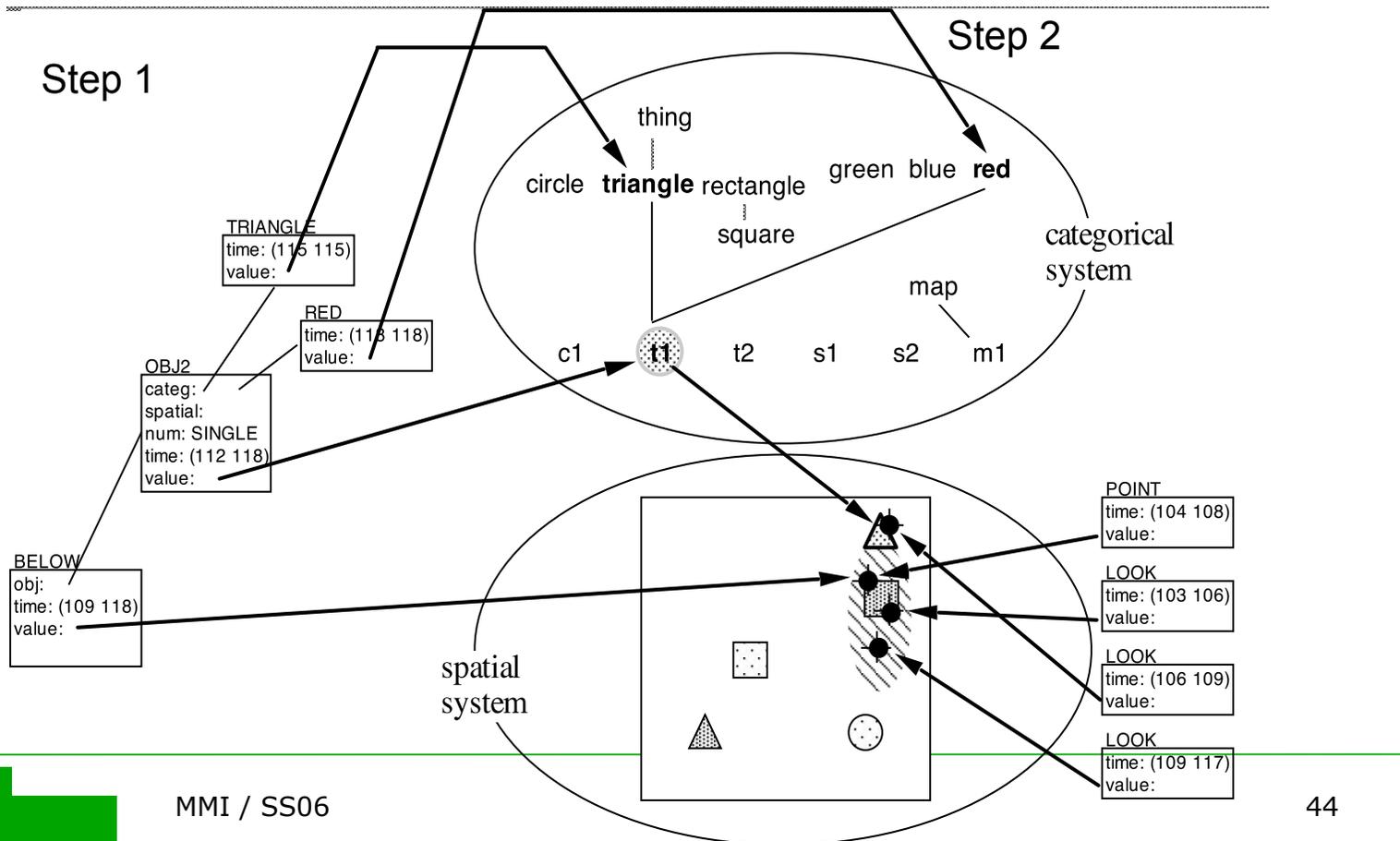
- Knowledge base spans two interconnected representational systems, objects are represented in both
  - categorical system (semantic network)
  - spatial system (locations)



# ICONIC: Evaluation

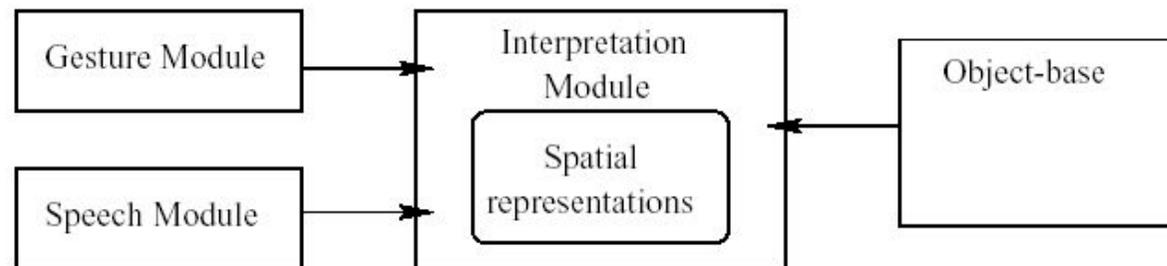
“...below the red triangle”

- Finds values for each frame in space/category systems
- Integrates spatial values from speech, gesture, eye



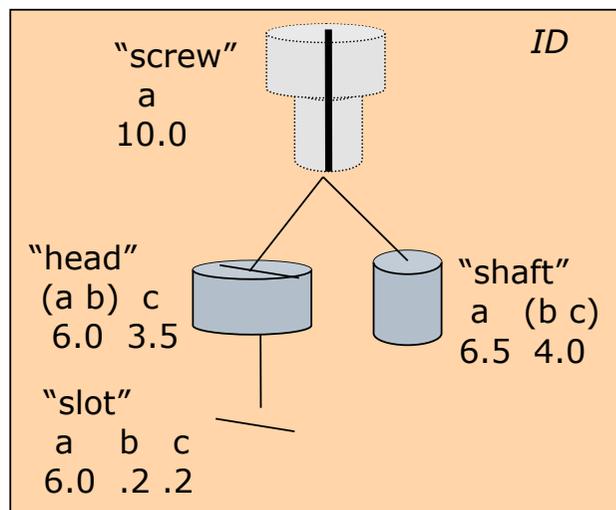
# ICONIC System

- Processes speech and gesture parallel (deictic, iconic and panomimic gestures)
- **Speech-driven:**
  - Whenever speech suggests possibility of a gesture (e.g. “that”, “like this”), system looks for appropriate segment
  - Tested on ambiguity of language, whether gestures can dissolve this.
- *iconic mapping*
  - Iconicity evaluated for hand posture, gestlets transformed to shape description
  - correlated to shape of objects referenced verbally

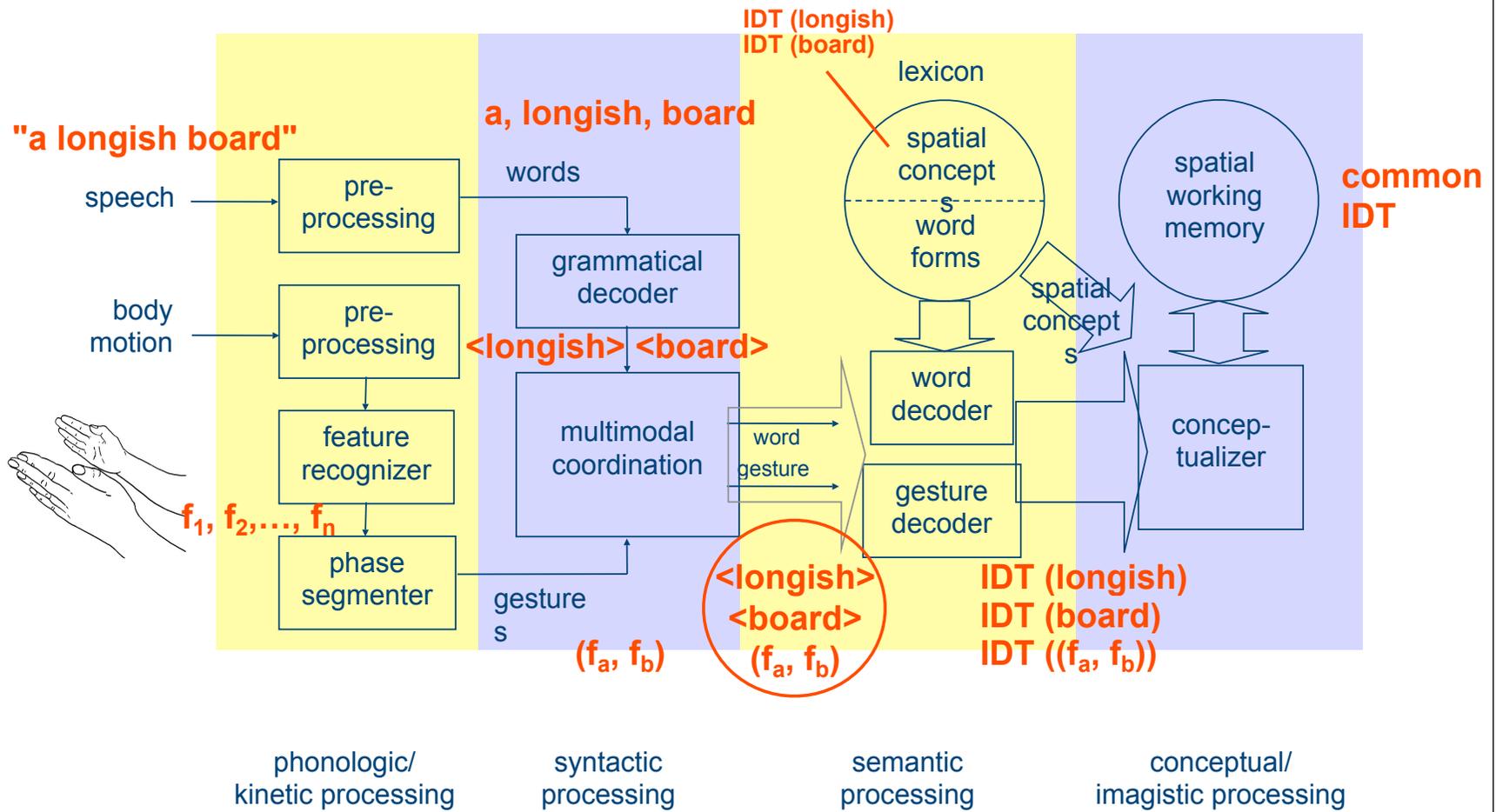


# Shape-related expressions (Sowa)

- translate gesture features into spatial representation of shape
- not limited to a single gesture, properties may accumulate over a series of movements and postures
- match shape representation with system's representation of how the objects look like



# Model overview



# Multimodale Ausgabe (fission)

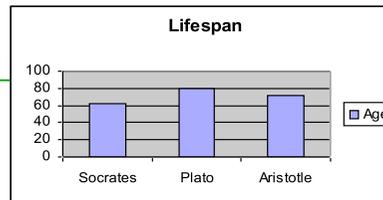
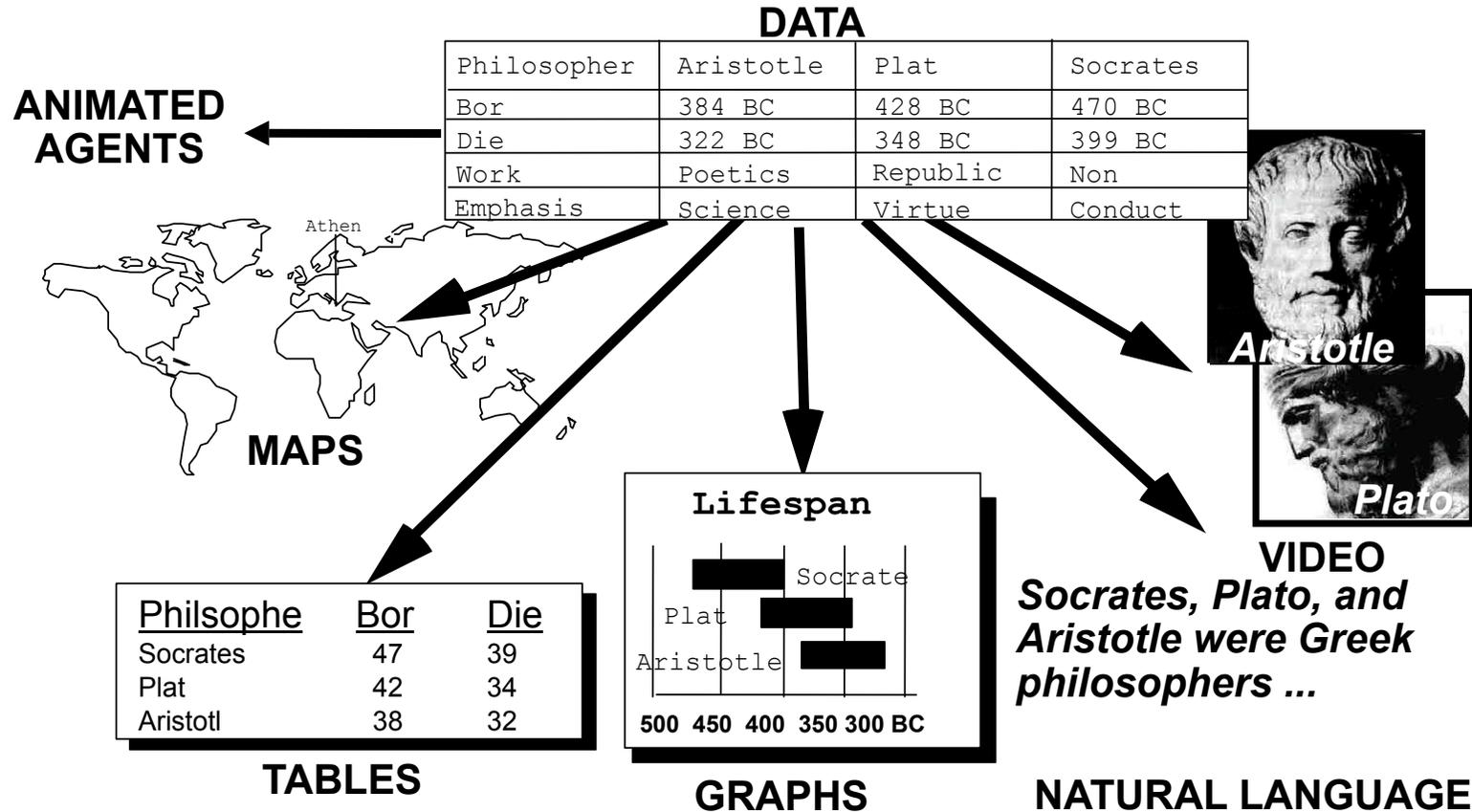
Zwei Ansätze zur Generierung multimodaler Ausgaben:

- "Klassische" Multimediasysteme:  
Informationen können über verschiedene Modalitäten oder eine Kombination von Modalitäten präsentiert werden. Es kommen die "üblichen" vom Desktop-Rechner gewohnten Modalitäten zum Einsatz: *Text, Grafik, Animation, Ton*
  
- Anthropomorpher Ansatz:  
Das System wird durch eine humanoide Figur als Kommunikationspartner verkörpert, oder dem System wird eine Figur als "Helfer" zur Seite gestellt.  
Hier stehen die natürlichen Modalitäten im Vordergrund:  
*Sprache, Gestik, Mimik, Körperhaltung*

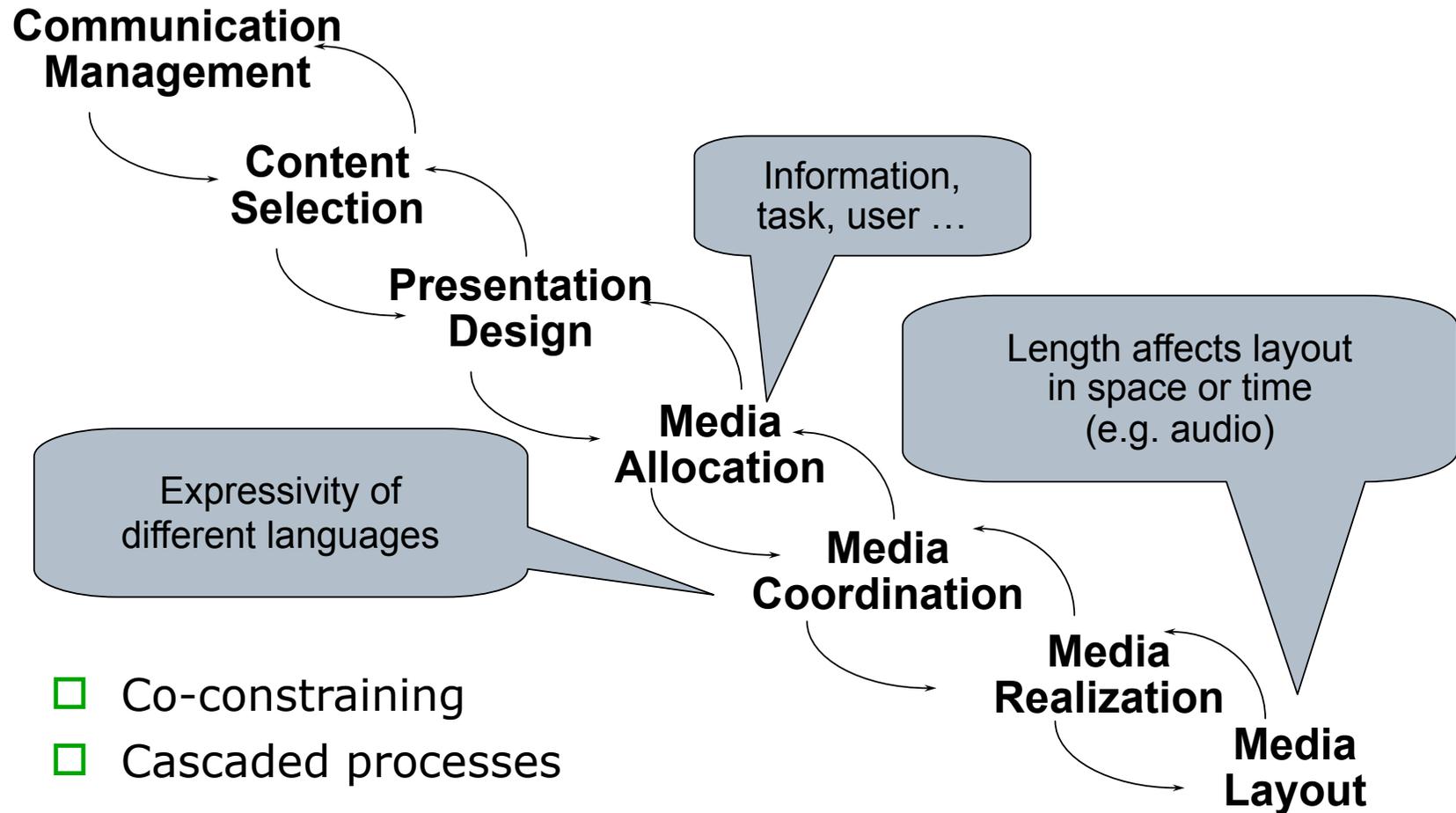


# Multimedia Presentation Generation

"No Presentation without Representation"

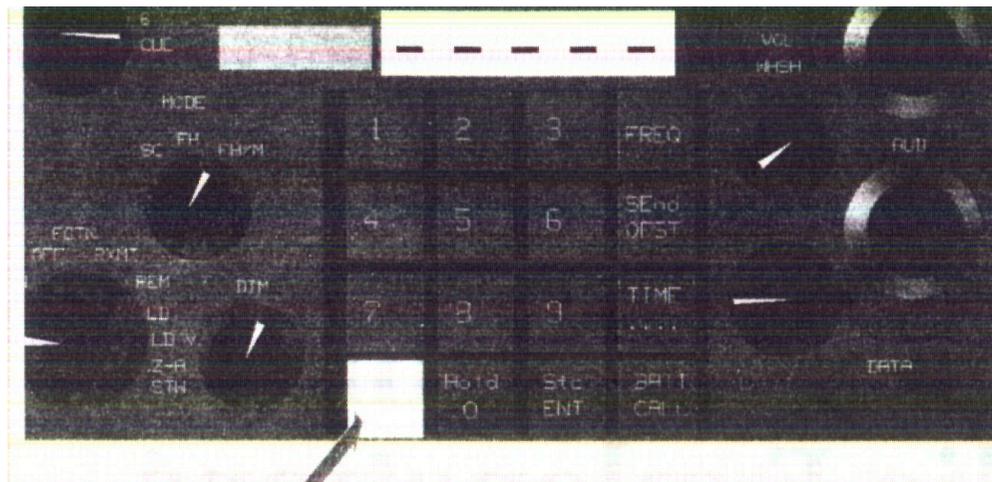


# Common Presentation Design Tasks



## COMET (**C**oordinated **M**ultimedia **E**xplanation **T**estbed; Feiner et al. 1993)

- System erklärt den Diagnosevorgang bei technischen Problemen mit Funkgeräten
- erst Bestimmung *was* das System ausdrücken will, dann Planung *wie* dieser Inhalt möglichst passend übermittelt wird



Press the CLR button to clear the display

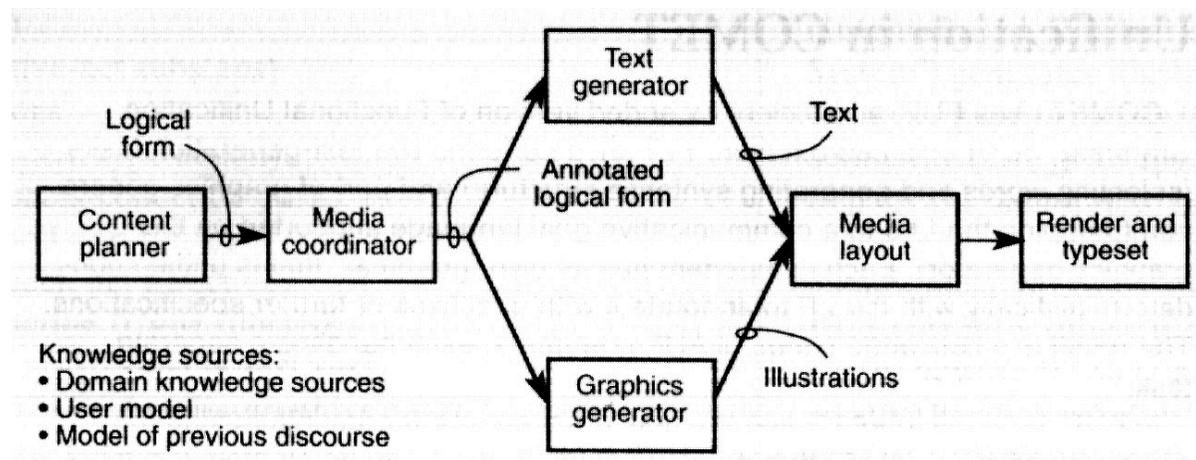


# Media coordination in COMET

Entscheidet heuristisch, welche Information in welcher Modalität dargestellt wird

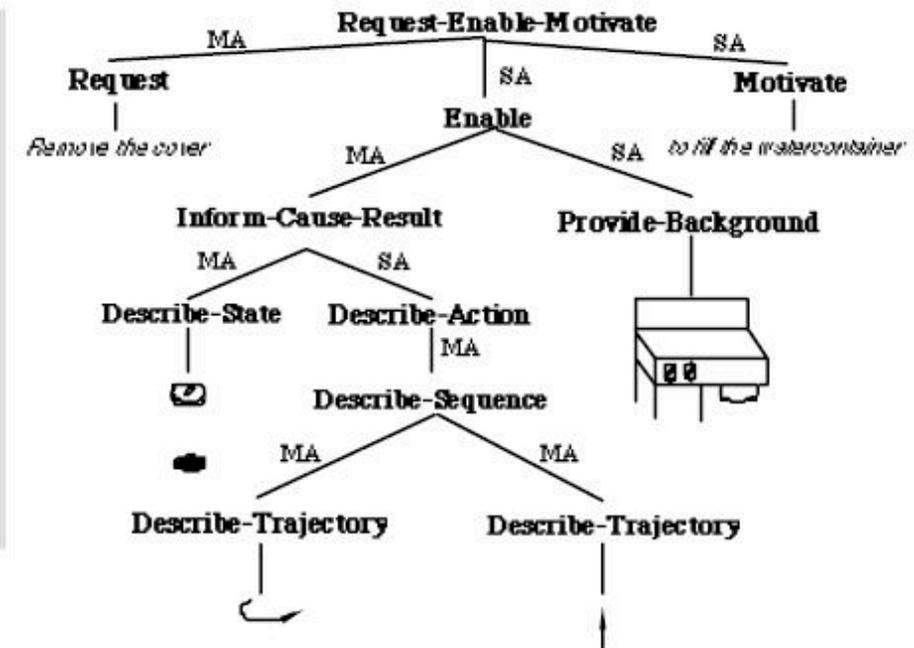
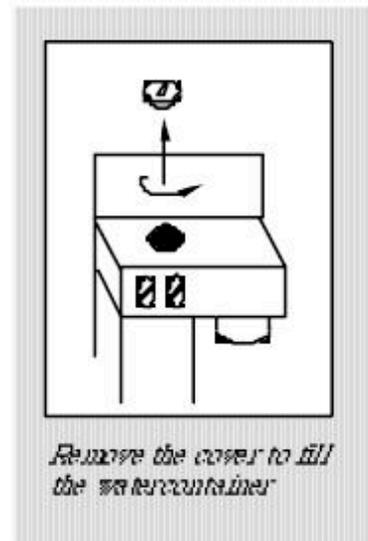
□ Klassifikation der Informationen in 6 Informationstypen in der ‚logical form‘ für Text vs. Grafik:

- Lokation, physische Attribute (Form etc.) → nur Grafik
- abstrakte Aktionen und Zusammenhänge zwischen Aktionen (Reihenfolge, Kausalität) → nur Text
- konkrete Aktionen → Grafik + Text



# WIP: Use of communicative acts

- Integrated planning process to create document plan
- Use of repository of **communicative acts** (cf. speech acts)
- Goal-refinement into subgoals
  - communicative (e.g., describe)
  - textual (e.g., S-request)
  - graphical (e.g., depict)



Wahlster et al., 1993;  
Andre & Rist, 1993



# Agent-based interfaces

- Natürliche Interaktion mit „natürlichem“ Gegenüber

