Human-computer interaction

Multimodal 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 <*d*,*L*> eines Interaktionsgeräts *d* mit einer Interaktionssprache *L*

Modalität & multimodal

Im folgenden verwendete Definition:

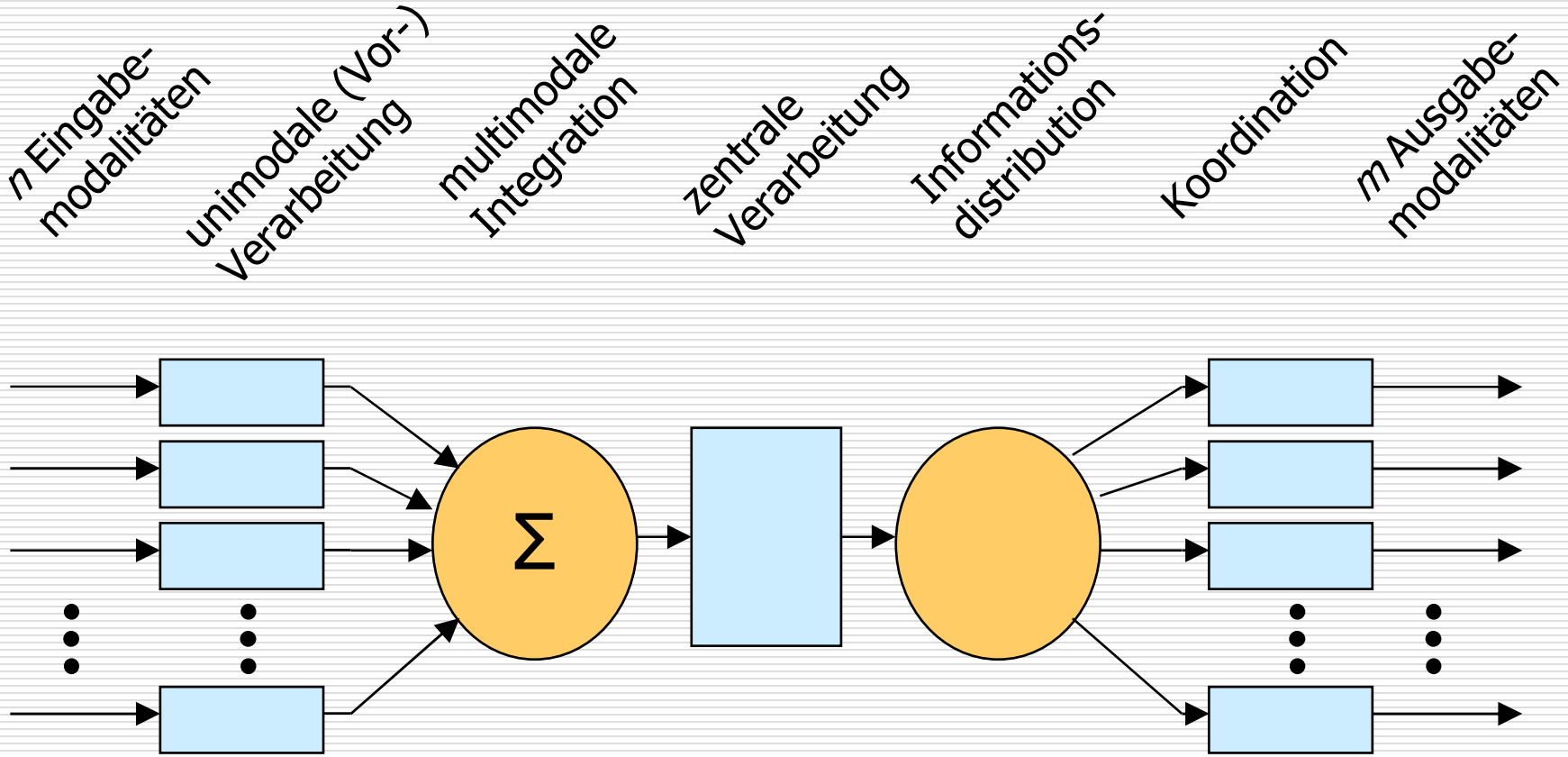
Eine Modalität bezeichnet ein kommunikatives System, das durch die Art und Weise wie Information codiert und interpretiert wird, gekennzeichnet ist.

- 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 Eingabe-modalitäten 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)*
- Die natürlichen Modalitäten sind kulturabhängig! Ausnahme: Emotionen, die durch Mimik ausgedrückt werden
- Neben den natürlich vorhandenen können Modalitäten erlernt werden: Kulturtechniken wie z.B. Lesen & Schreiben oder auch *point-and-click* als "Subkulturtechnik"

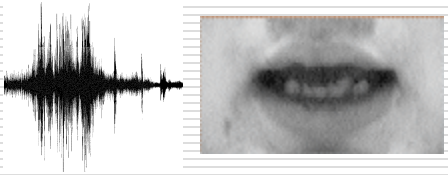
Multimodale Systeme: Prinzip

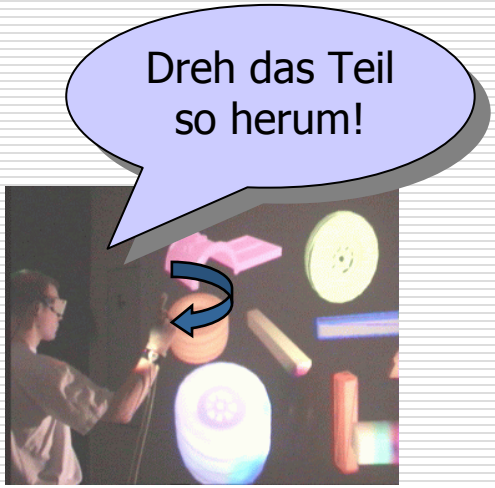


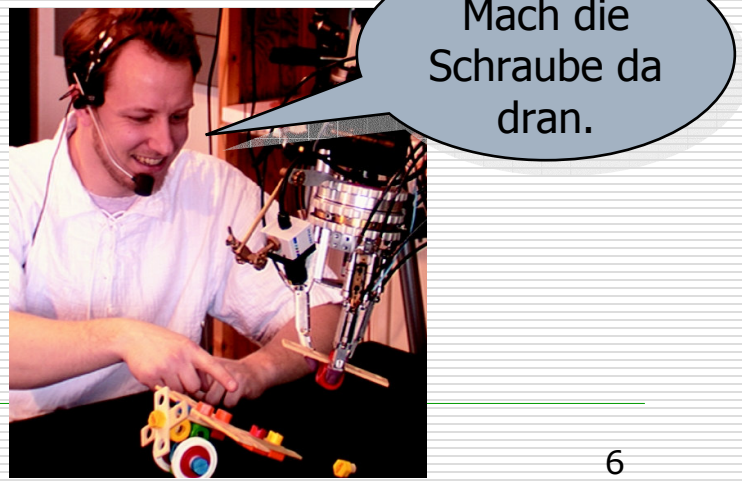
Verarbeitungsschema für multimodale Systeme

Multimodale Systeme: Beispiele

Spracherkennung
+ Lippenlesen







Warum multimodal?

Natürlichkeit & Intuitivität

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

Bandbreite & Effizienz der Informationscodierung

- Kommunikation von *mehr* Information pro Zeit möglich

Adäquatheit der Informationscodierung

- unterschiedliche Informationsarten lassen sich in verschiedenen Modalitäten verschieden gut übermitteln

Alternative Kommunikationswege *(universal design)*

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

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

Multimodal Interfaces vs. GUIs

GUIs

1. Assume that there is a single event stream that controls event loop with processing being sequential
2. Assume interface actions (e.g. selection of items) are atomic and unambiguous
3. Built to be separable from application software and reside centrally on one machine
4. Do not require 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. Have large computational and memory requirements and are typically distributed over the network (often as multi-agent systems)
4. Require time stamping of input and development of temporal constraints on mode fusion operations

Natural input/output modalities

Speech
Gesture
Mimics
Eye gaze
Body language & proxemic

Sprache

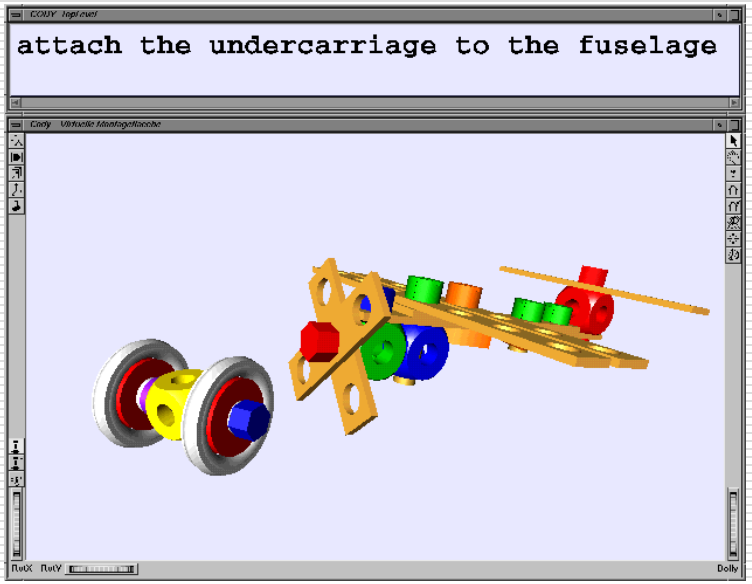
- *symbolische* Modalität: sprachliche Zeichen bezeichnen nur aufgrund von Konventionen
  - Ausnahme: *Onomatopoetika* (Lautmalerei)
- (gesprochene) Sprache umfaßt weitere nichtsymbolische Information: Prosodie

(in der VL bereits ausführlich behandelt)

Speech: *Task-level* communication

(Zeltzer, 91; Zeltzer & Gaffron, 96)

- convey task to be done, choose desired mode of manual operation, interface correction channel
- instruct the system in an intuitive, efficient, and robust way

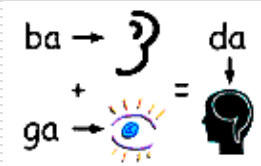


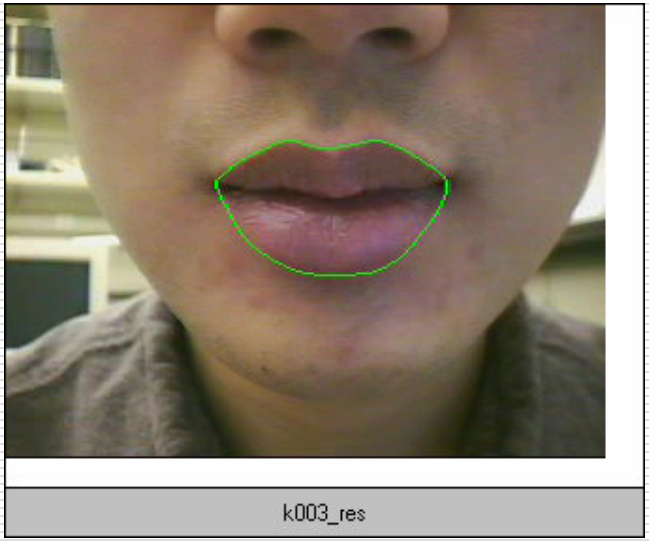
Communication is situated

- referring to objects in terms of their actual or future *roles* (a screw is an axis, a bar is a propeller, …)
- referring to current spatial properties and relations (relative location, orientation, shape)

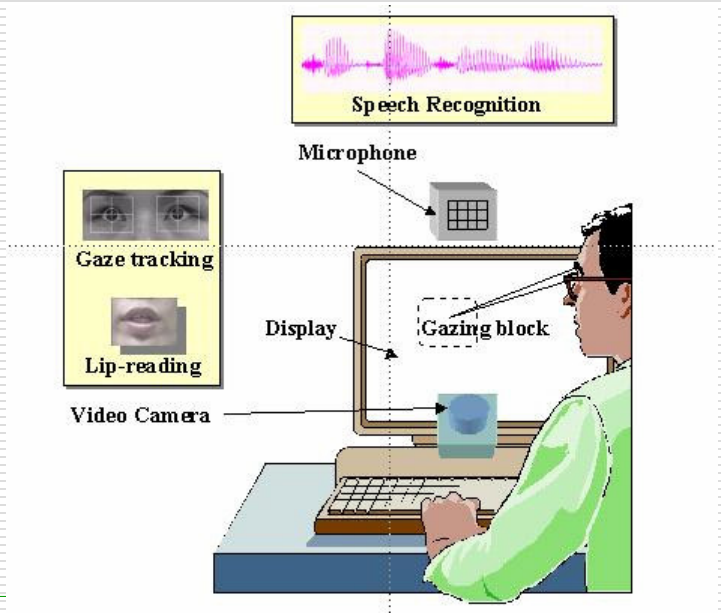
Sprechen: Lippenlesen

- 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: Klassen

Es gibt viele Klassifikationsschemata. Weitere Gestentypen:

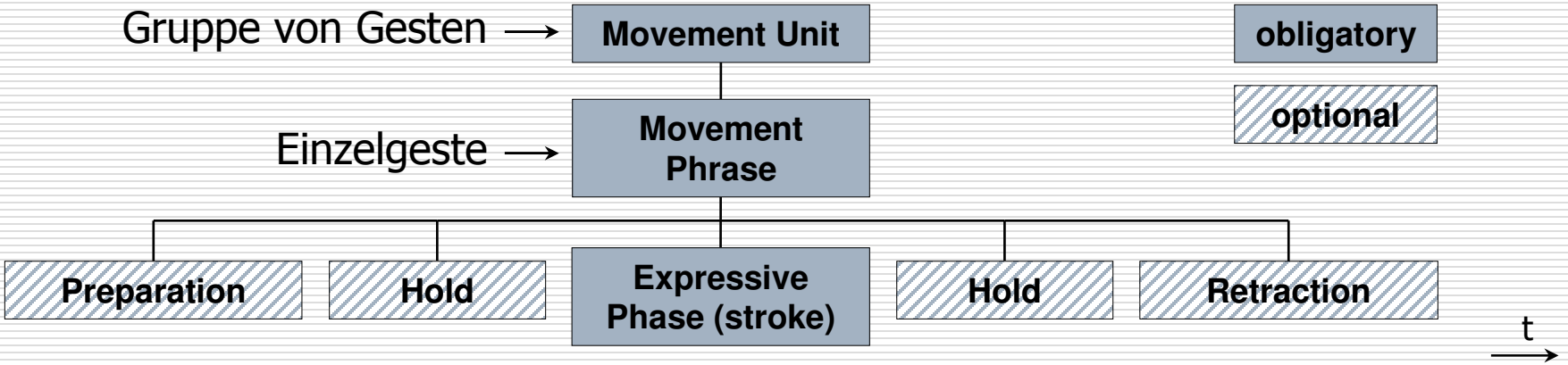
- Beats / Batons

  (häufig) zweiphasige kurze Schlag-Gesten, die den Rhythmus der Sprache sowie Betonungen markieren
- metaphorische Gesten

  wie ikonische, nur stellen sie eine Metapher bildlich dar
- Weitere Subtypen ikonischer Gesten
  - spatiographisch: räumliche Konfigurationen
  - kinemimisch (pantomimisch): Aktionen
  - pictomimisch: räumliche Eigenschaften (Größe, Form,…)

Gesten: Struktur

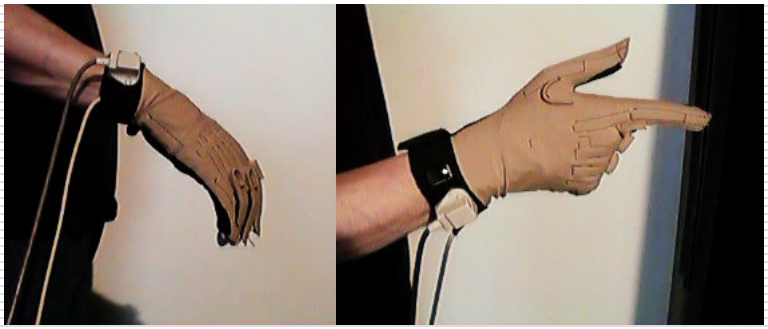
Gesten bestehen typischerweise aus mehreren Bewegungsphasen

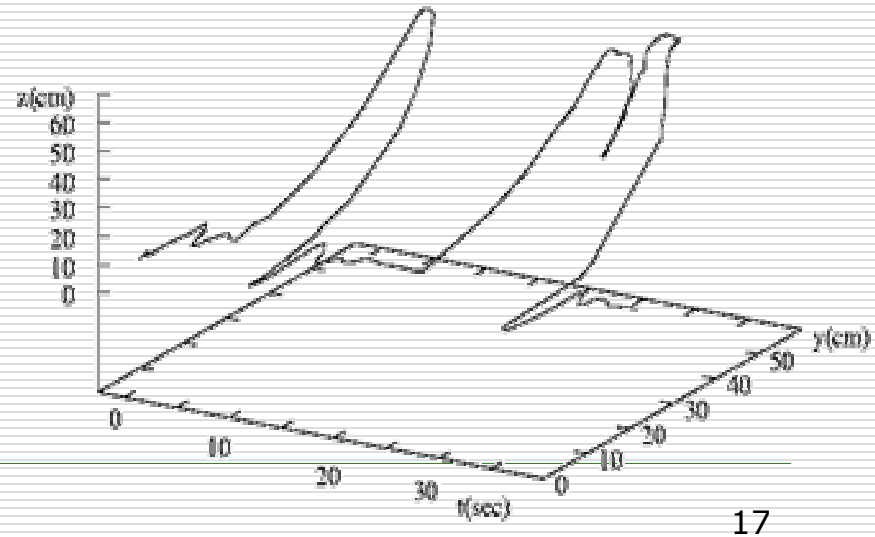


- Vorbereitungsphase (preparation): Hände in Stellung bringen
- expressive Phase (stroke): bedeutungstragender Teil
- Rückführungsphase (retraction): Hände zurück in eine Ruhestellung
- Haltephasen (hold): keine Bewegung

Gestenerkennung

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





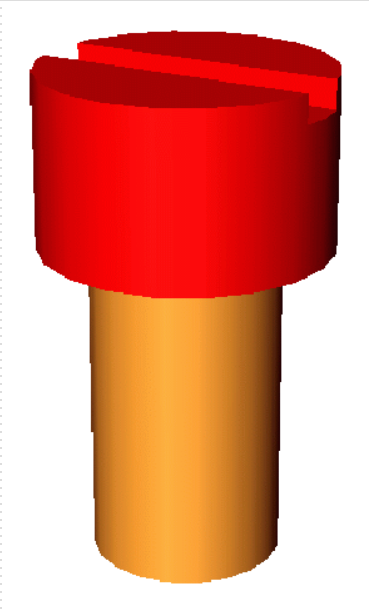
Natürliche Multimodalität: Gestik und Sprache

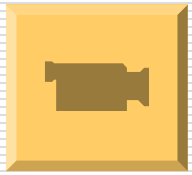
Zwischen Sprache und Gestik besteht ein zeitlicher und inhaltlicher Zusammenhang – zusammengefaßt in drei "Regeln" (McNeill, 1992):

- 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.

Beispiel

Beschreibe das Objekt!



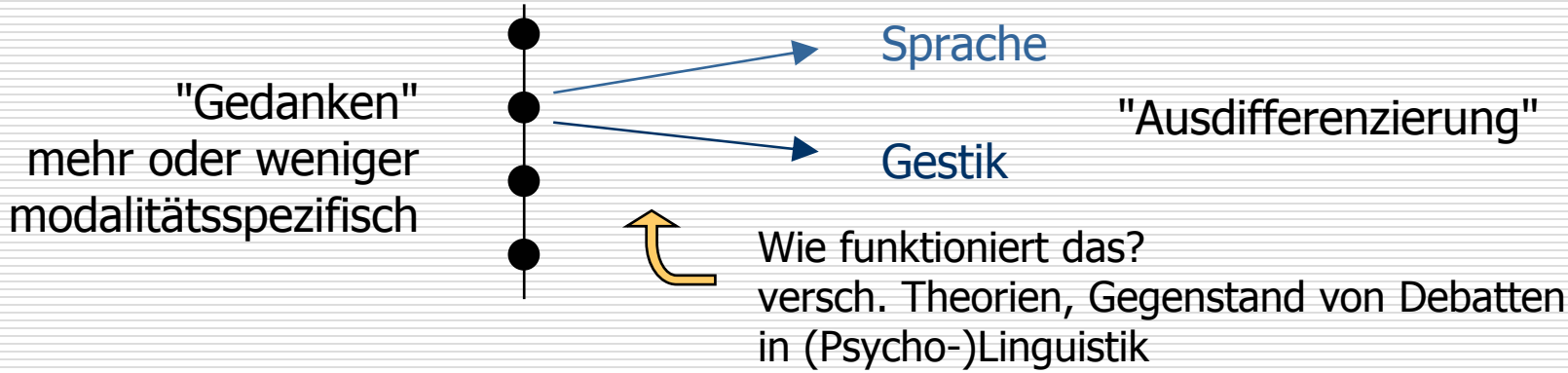


Gestik und Sprache

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

Vorstellung

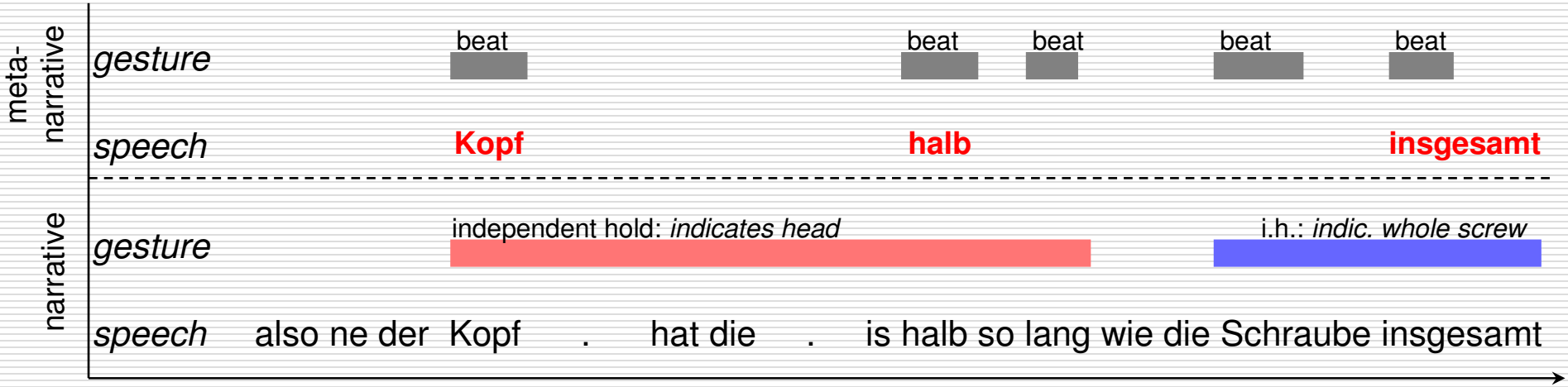
Kommunikation: Abfolge von zu kommunizierenden Inhalten oder Gedanken. Gedanken entwickeln sich dann zu (oder werden verpackt in) Sprache und Gestik.



Ist Gestik *eine* Modalität?

Überlagerung

- ikonische Gesten und Beats repräsentieren zwei Inhaltsstränge (*narrativ* and *meta-narrativ)*
- Stränge können sich überlagern



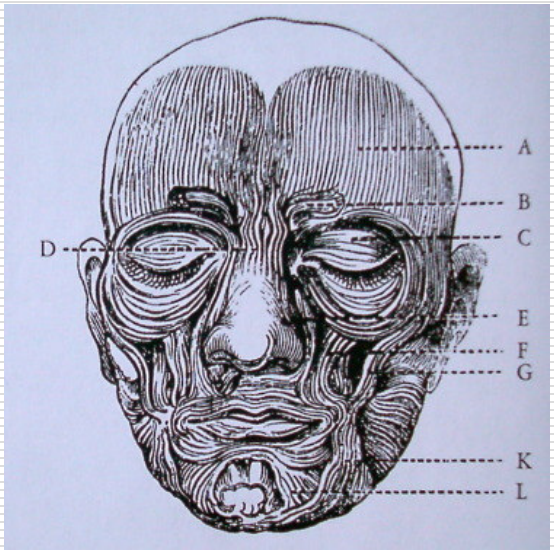
Mimik

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 wesentlich zum Ausdruck *emotionaler Zustände* 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

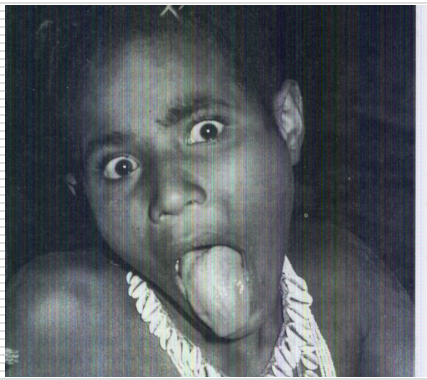
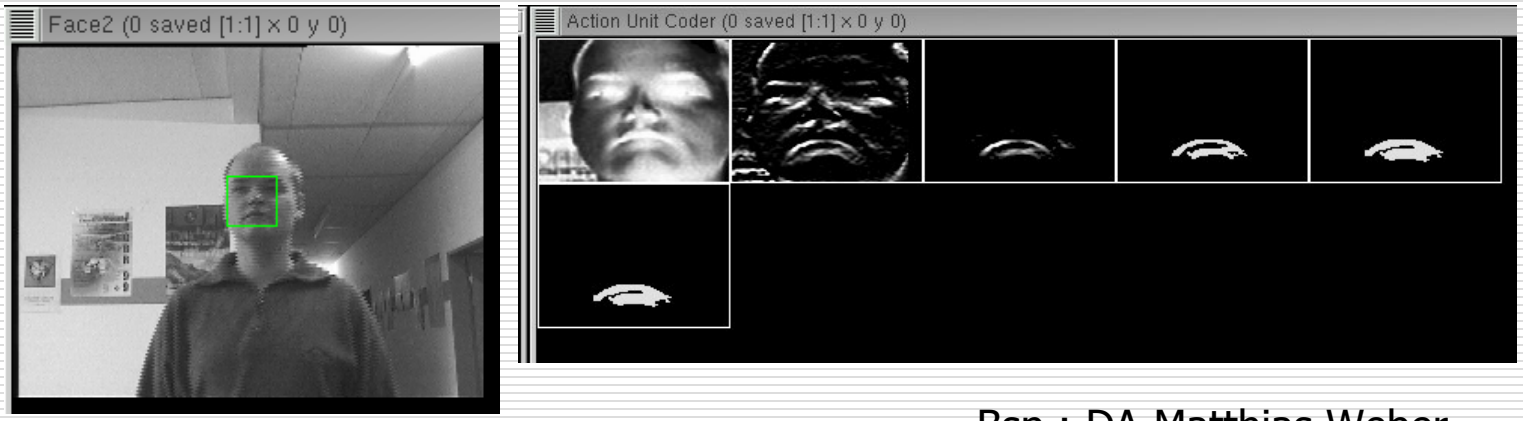


Abb.: Schiefenhövel et al., 1994

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

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…"

Körpersprache & Proxemik

Körpersprache

Art und Weise wie die Körperhaltung, Position und Orientierung zur Kommunikation genutzt wird

Lexikondefinition Proxemik (Merriam-Webster)

"the study of the nature, degree, and effect of the spatial separation individuals naturally maintain (as in various social and interpersonal situations) and of how this separation relates to environmental and cultural factors"

Edward Hall's Theory of Proxemics

People maintain differing degrees of personal distance depending on the social setting and their cultural backgrounds.

Multimodal analysis

Multimodal analysis

- 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

Trends…

- Earliest systems: supported speech input along with keyboard or mouse GUI interfaces.
- In ’80s and ’90s: systems were developed to use spoken input as an alternative to text via keyboard, e.g. CUBRICON, XTRA, Galaxy, Shoptalk and others.
- Late 90’s up to now: systems that allow to augment NL input with mouse/pen pointing, e.g. Quickset, etc.
- Recent system: designs based on two or even more parallel input steams both capable of conveying rich semantic information, e.g. speech and gesture, speech and gaze, etc.

Multimodal analysis

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 deute 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"*)

Terminology

Mutual disambiguation

disambiguation of signal or semantic-level information in one error-prone input mode from partial information supplied by another, with the net effect of suppressing errors experienced by the user.

Feature-level fusion

fusing low-level feature information from parallel input signals, applied to closely synchronized input such as speech and lip movements.

Semantic-level fusion

integrating semantic information derived from parallel input modes, used for processing speech and gesture input.

Frame-based integration

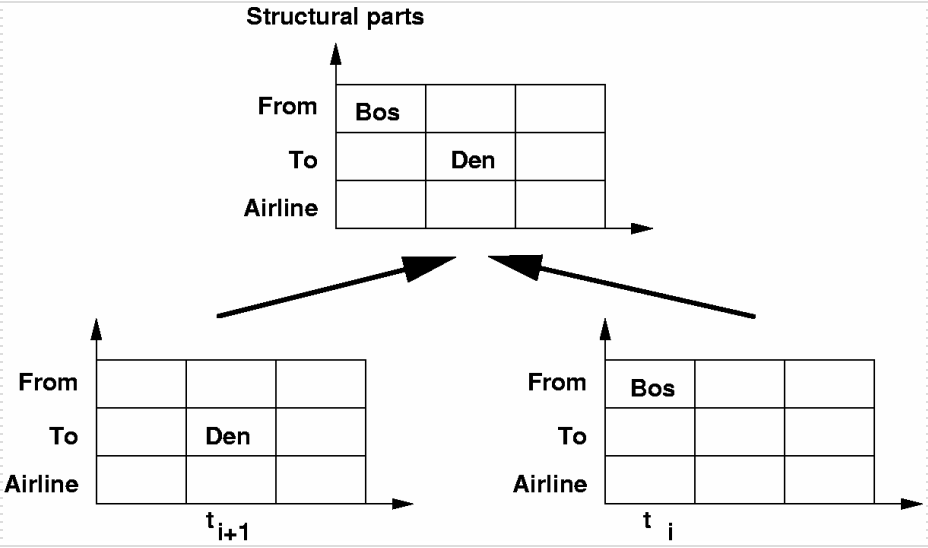
pattern matching technique for merging attribute-value data structures to fuse semantic information derived from two input modes into a common meaning representation

Unification-based integration

logic-based method for integrating partial meaning fragments derived from two input modes into a common meaning representation, derives from logic programming

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



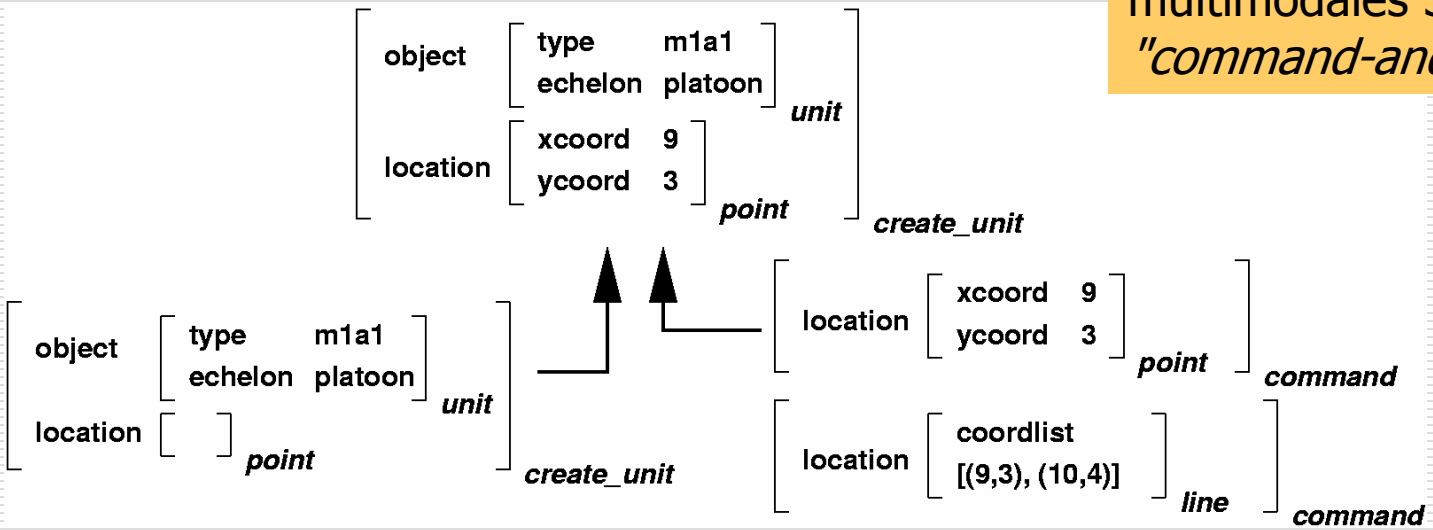
Beispiel: **MATIS**, **M**ultimodal **A**ir **T**ravel **I**nformation **S**ystem

(„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"*



from speech

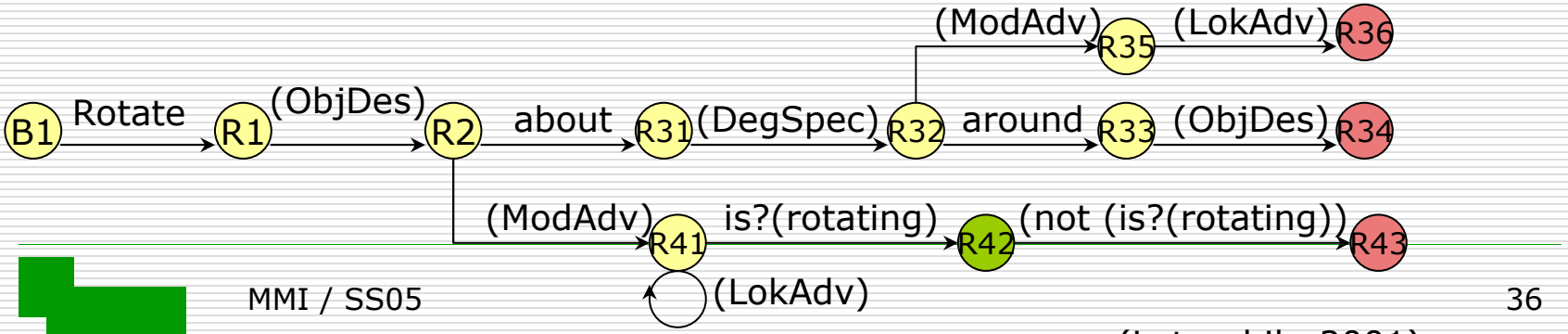
from pen/map

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

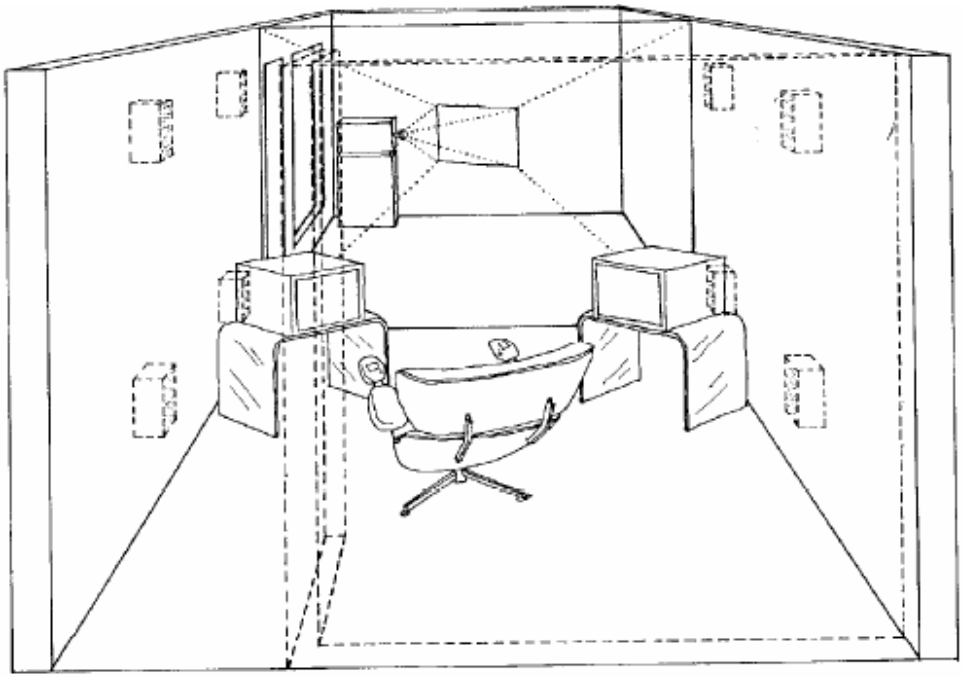
Beispiel: **tATN**

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



(Latoschik, 2001)

MIT Media Room



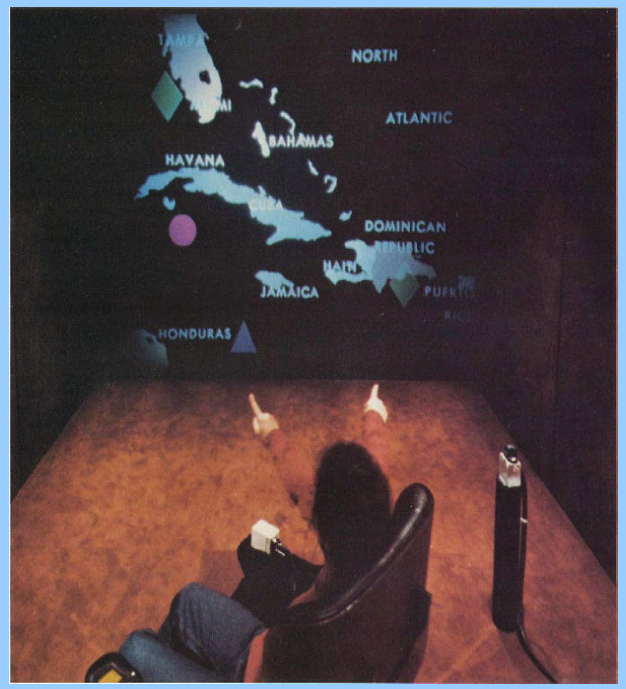
- loudspeakers, frosted glass projection screen, TV monitors on either side of user's chair
- chair arms with one-inch high joystick sensitive to pressure and direction, touch sensitive pad
- Position-sensing cube attached to wristband

Put-That-There (Bolt, 1980)

- One of the earliest multimodal concept demonstration using speech and pointing
- Created by Architecture Machine Group at MIT
- Quote from Richard Bolt:

  "Even after 17 years, looking at the video of the demo, you sense something special when Chris, seated before our media room screen, raises his hand, points, and says " Put that (pointing to a blue triangle)…there (pointing to a spot above and to the left)," and lo, the triangle moves to where he told it to."

Commands in Put-That-There



(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"

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”

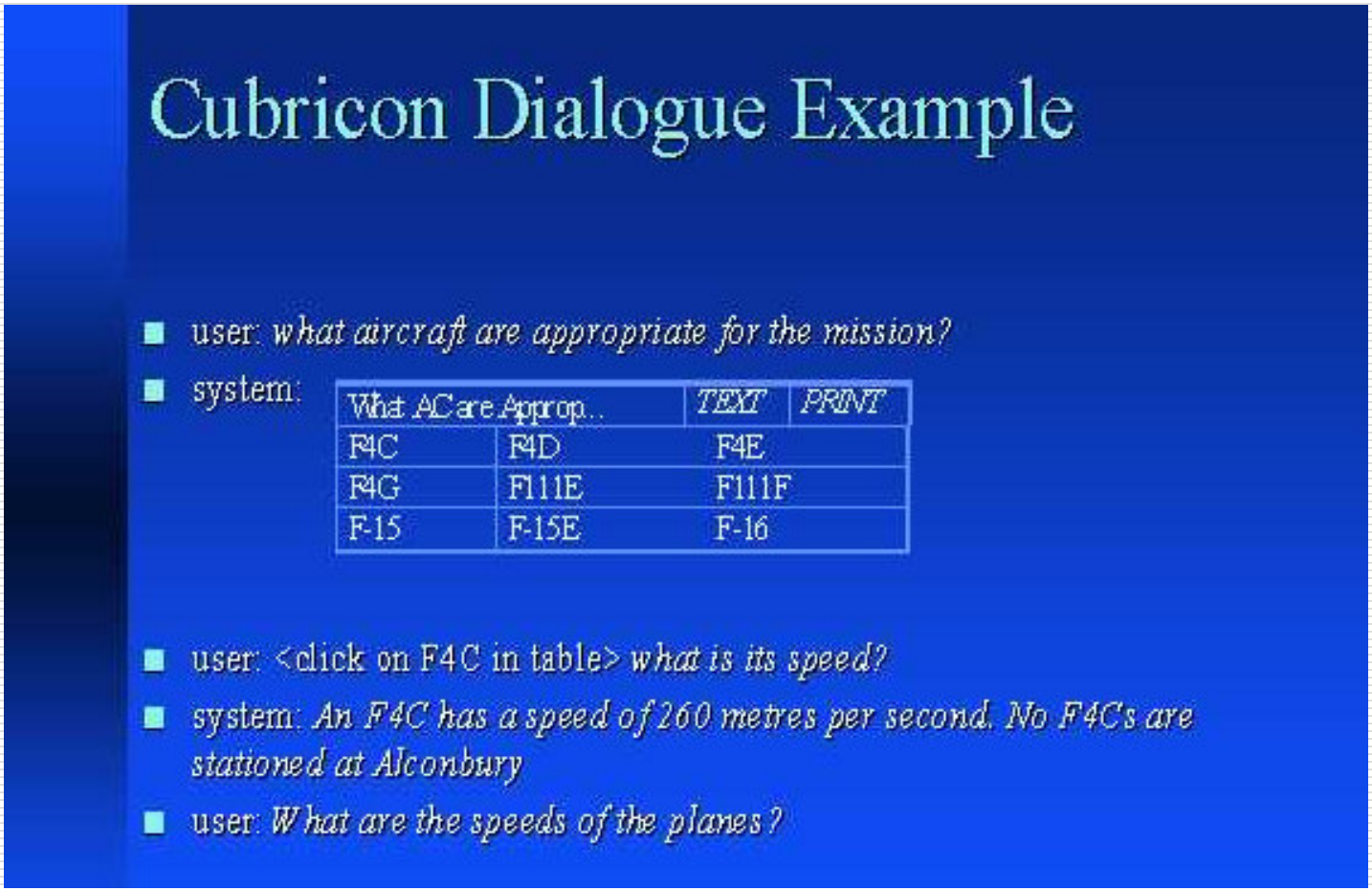
→ Recognizer sends code to host system indicating a naming command (“call”) → x,y coordinates of item signal are noted by host → host switches speech recognition to training mode to learn the (possibly new) name to be given to the object

All utterances processed with hard-wired procedural semantics

CUBRICON (Neal & Shapiro, 1991)

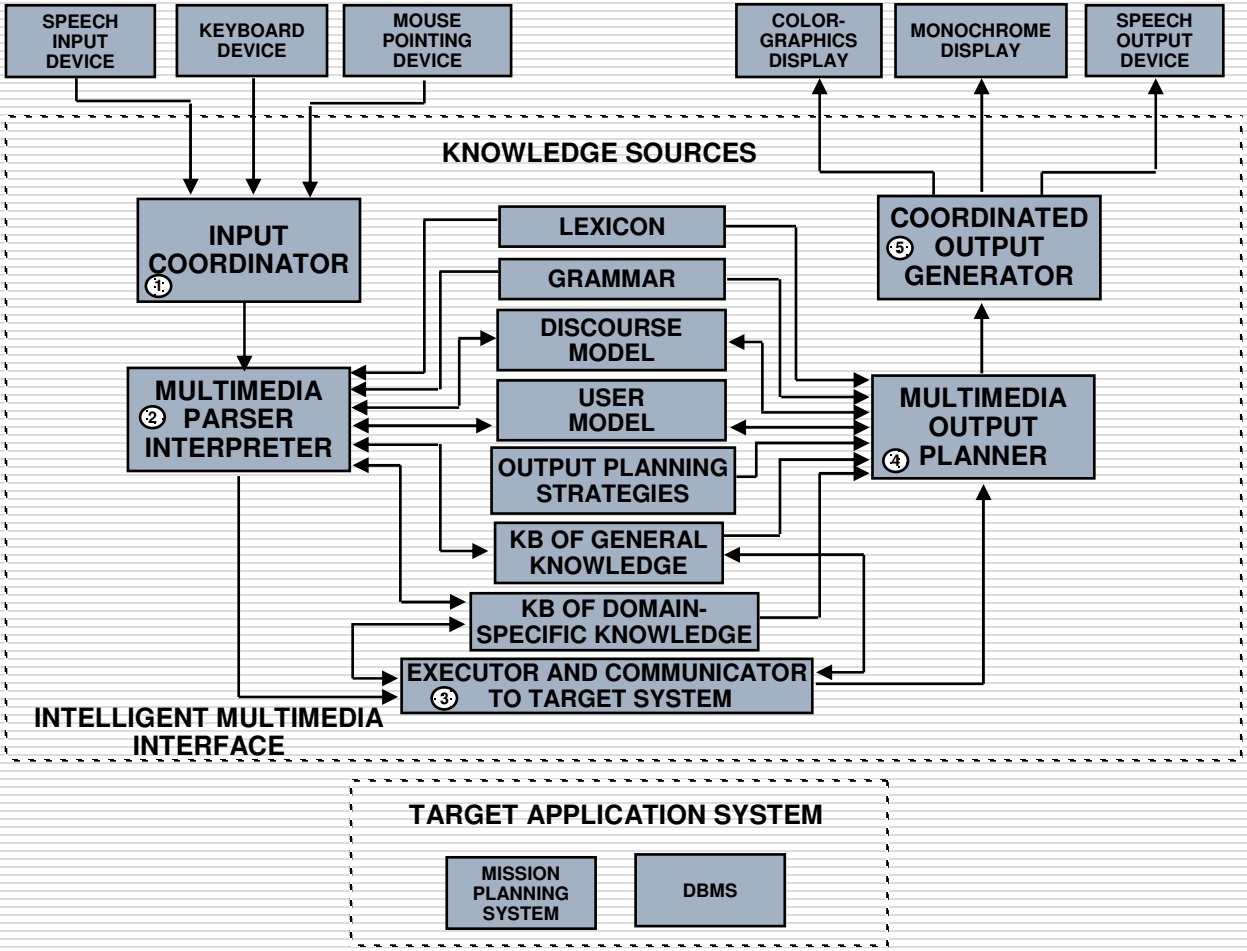
- System integrating deictic and graphic gestures with simultaneous NL for *both* user input and system output
- interface capabilities
  - Accepts and understands multimedia input – references to entities in NL can include pointing
  - Disambiguates unclear references and infers intended referent
 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

CUBRICON: example



Calspan-UB Research Center Intelligent CONversationalist

CUBRICON: Architecture



CUBRICON: Multimodal Language

- Spoken or written NL (NP & locative adverbials) and mouse gestures:
  - Variety in objects that can be pointed to: windows, form slots, table entries, icons, points.
  - Variety in number of point gestures allowed per phrase.
  - Variety in number of multimodal phrases allowed per sentence.
  - Interrogative: *"Is this <point> a SAM?"*
  - Imperative: *"Enter this <point-map-icon> here <point-form-slot>."*
  - Declarative: *„Units from this <point-1> airbase will strike these targets <point-2> <point-3> <point-4>."*
- Multimedia parser: ATN network for NL + mouse gesture

CUBRICON Knowledge Sources

- 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)

CUBRICON: referent determination

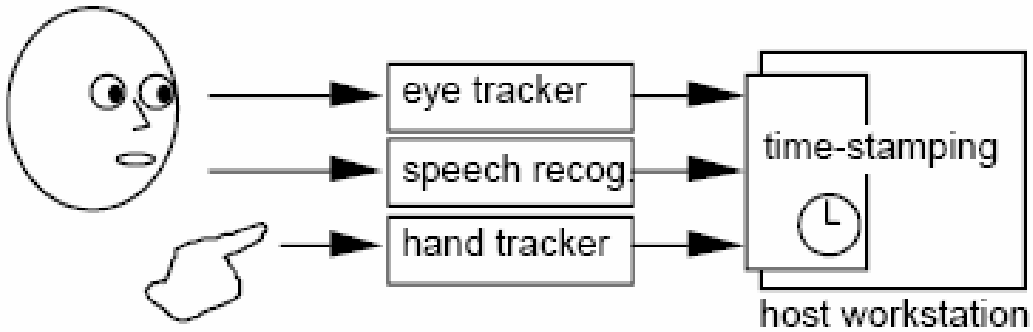
- Employs mutual disambiguation. Takes also care of pointing that is inconsistent with NL by using information from the sentence as filtering criteria for candidate objects.
  - Example:
    User: "*What is the mobility of these <point> <point>*?"
    System uses "*mobility*" property to select from candidate referents of the point gesture – uses display model and knowledge base

- Starting at display position indicated by point gesture, incremental bounded search is performed to find at least one object consistent with the semantic features expressed in speech

ICONIC (Koons et al., 1993)

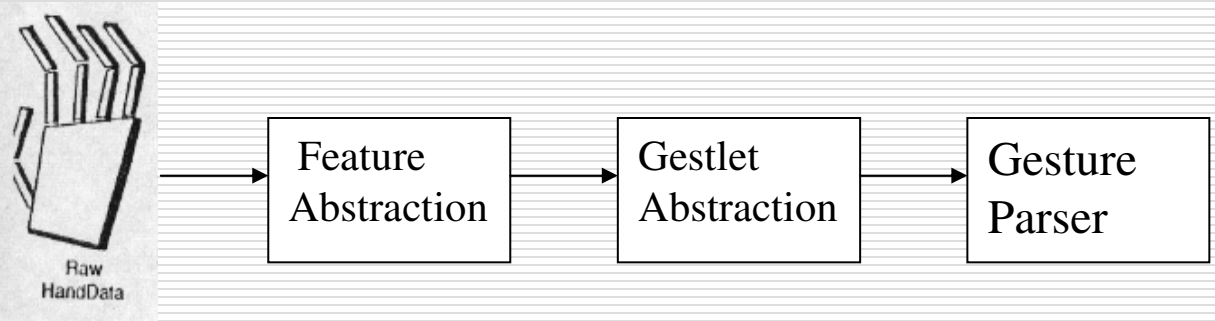
- Integrating simultaneous speech, gestural, and eye movement (for reference resolution for map and blocks world interaction)
- Problems: timing and abstraction
  - All three streams of data are collected on a central workstation and assigned time stamps, used later to realign data

Example: *"move the teapot like this"* + dynamic gesture indicates direction



ICONIC: Representing gaze & gesture

- Eye input: fixations, saccades, blinks
- Gesture input: Features
  - Posture: straight, relaxed, closed
  - Orientation (normal and longitudinal vectors from palm): up, down, left, right, forward, backward
  - Motion: moving, stopped
- Gestlets: Stream of gesture features
  - e.g., Pointing = attack, sweep, end reference



ICONIC: Processing input streams

Step 1 - Parsing

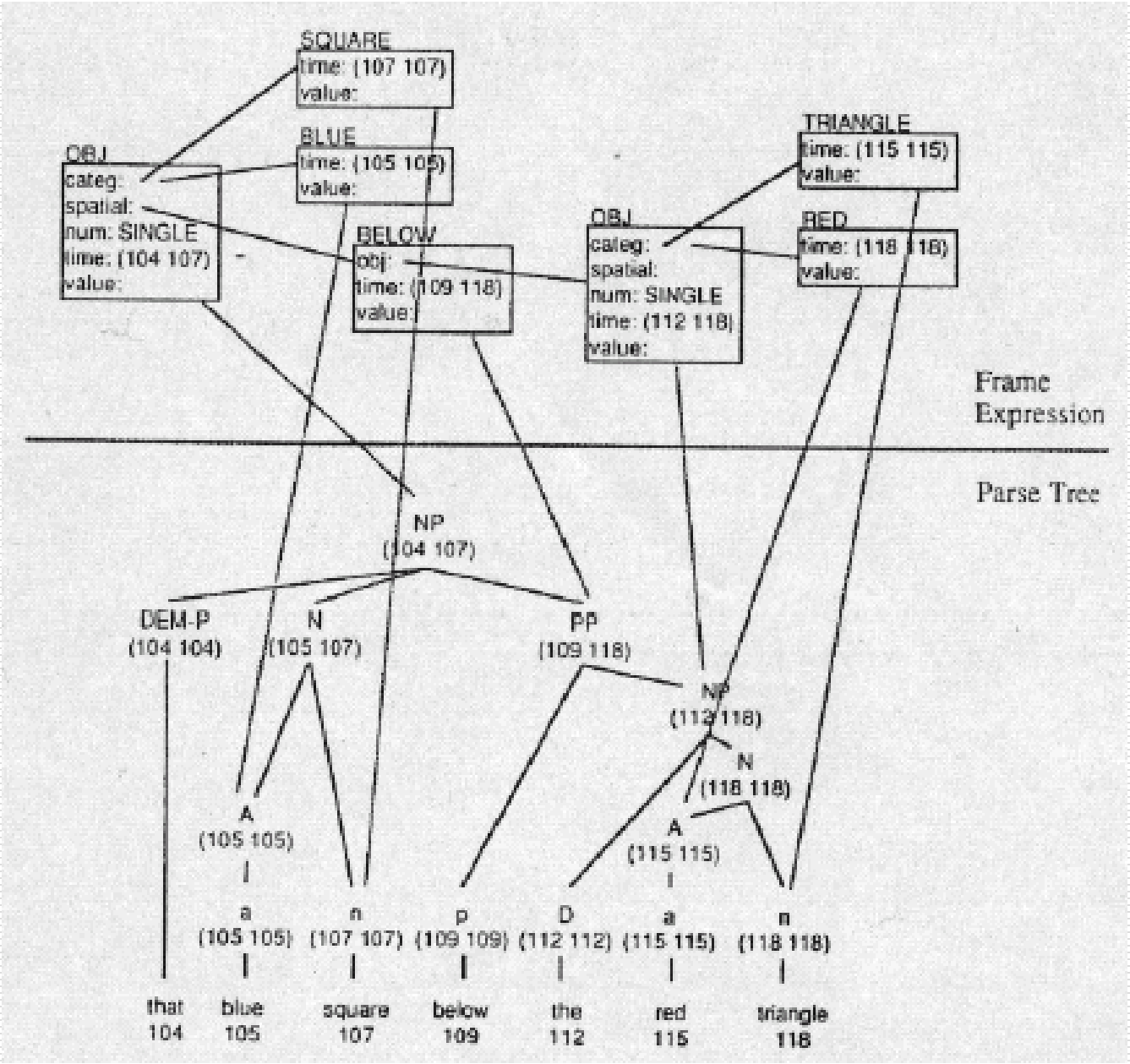
- Parse input data stream
- Generate frame-based description of the data

Step 2 - Evaluation

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

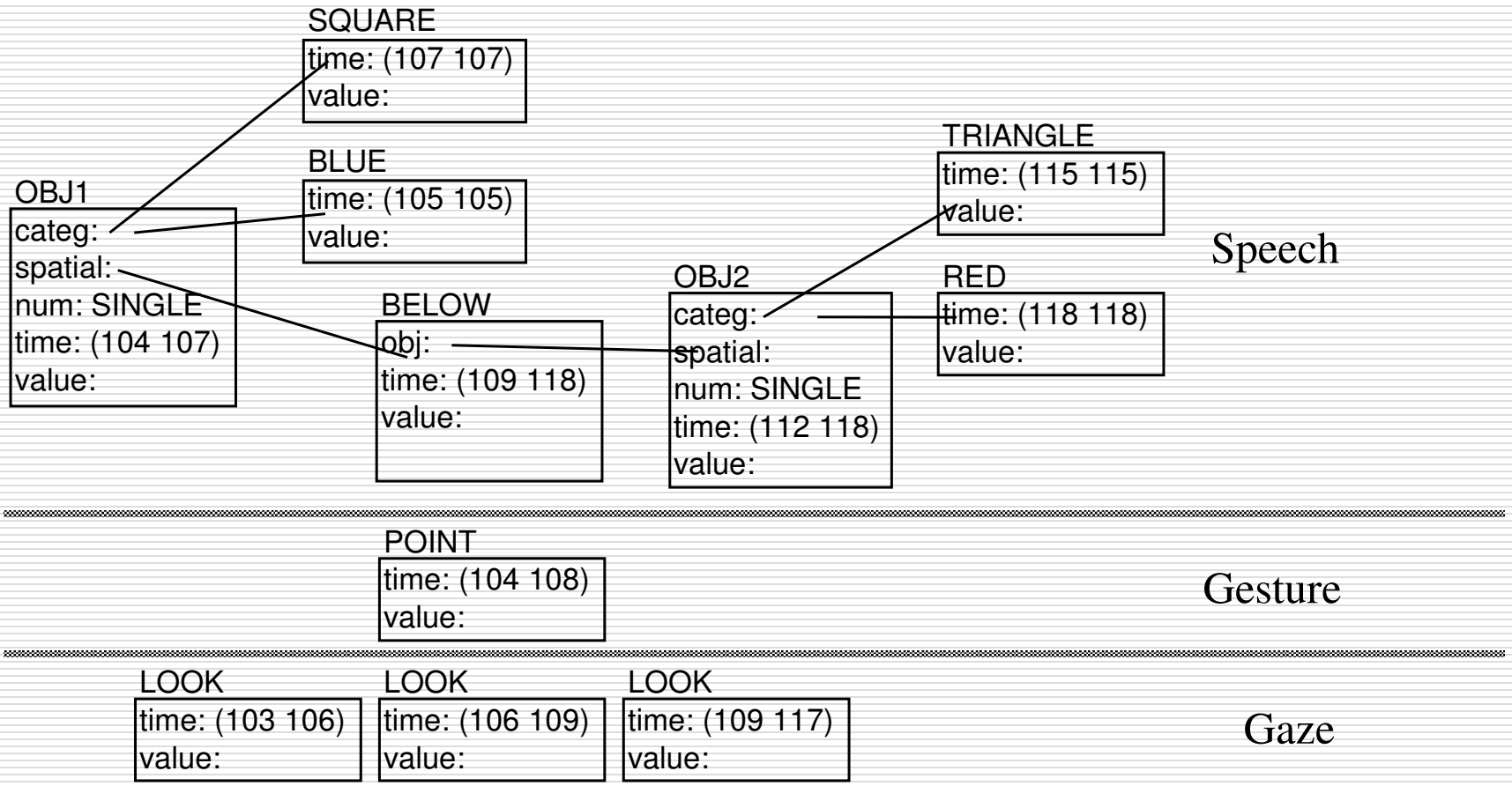
- Knowledge base spans two interconnected representational systems, objects are represented in both
  - categorical system (semantic network)
  - spatial system (locations)

ICONIC: Parsing



ICONIC: Frame-based integration

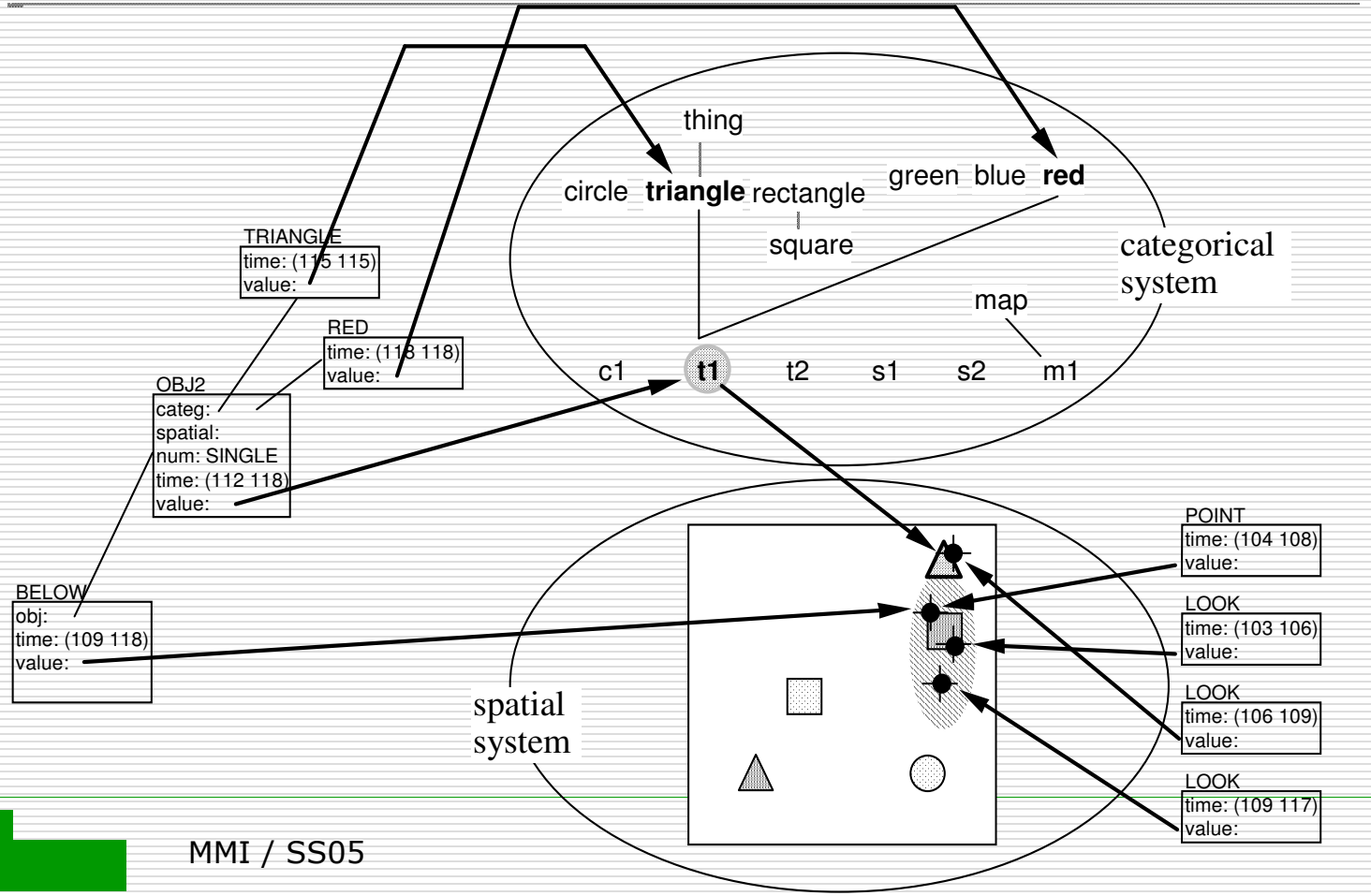
- Idealized example of frames produced during utterance
  " ... that blue square below the red triangle"



ICONIC: Evaluation

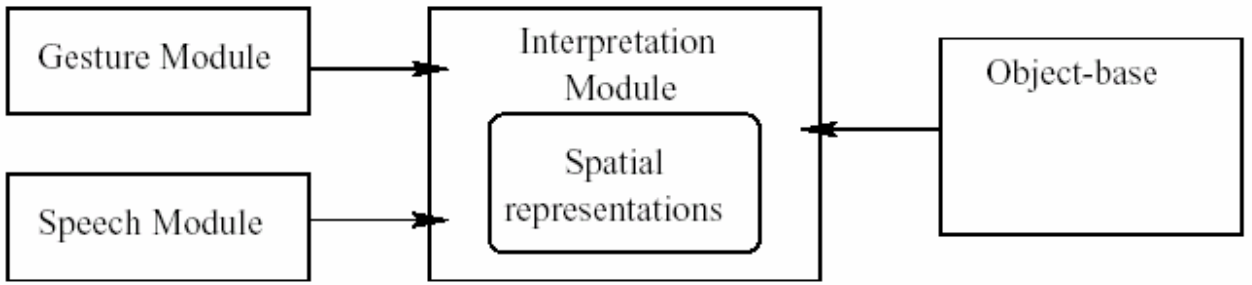
"…below the red triangle"

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



ICONIC System

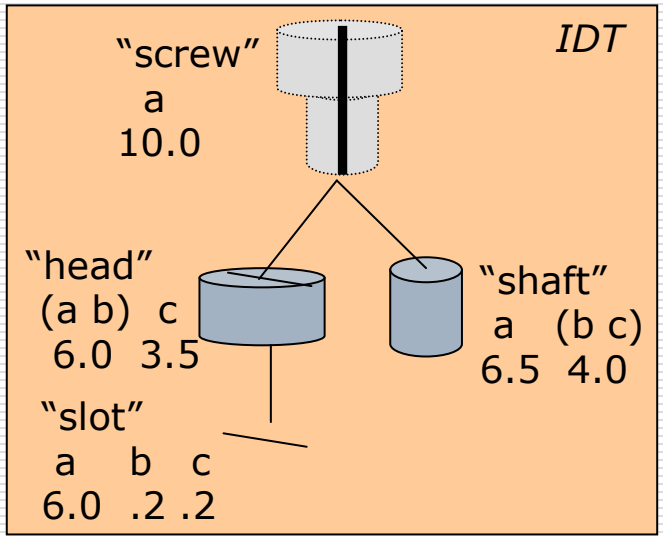
- Processes speech and gesture parallel (deictic, iconic and panomimic gestures)
- Speech-driven:
  - Whenever speech suggests possibility of a gesture (e.g. “that”, ”like this”), system looks for appropriate segment
  - Tested on ambiguity of language, whether gestures can dissolve this.
- *iconic mapping*

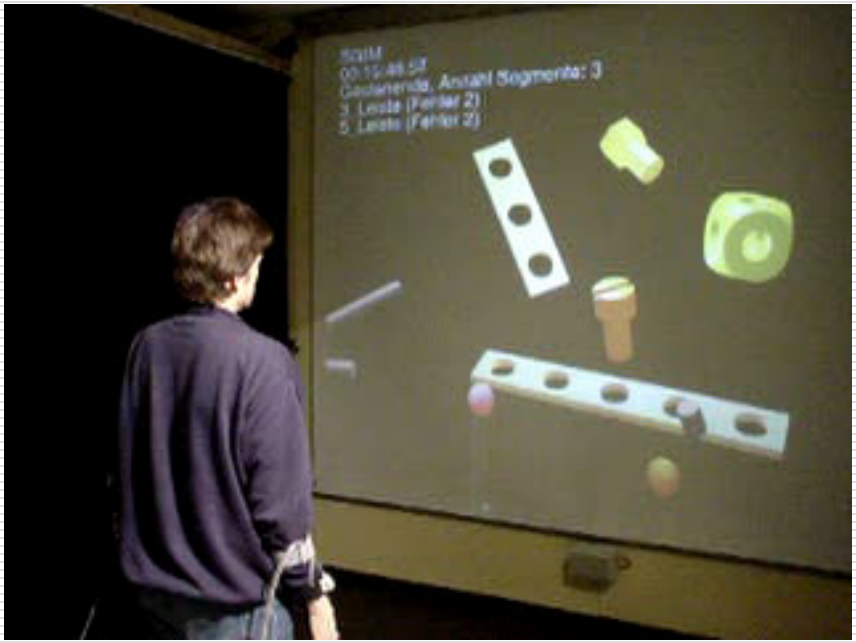


bally

Iconic gesture interpretation

- 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





(Sowa, 2005)

Multimodal generation

Multimodale Ausgabe (fission)

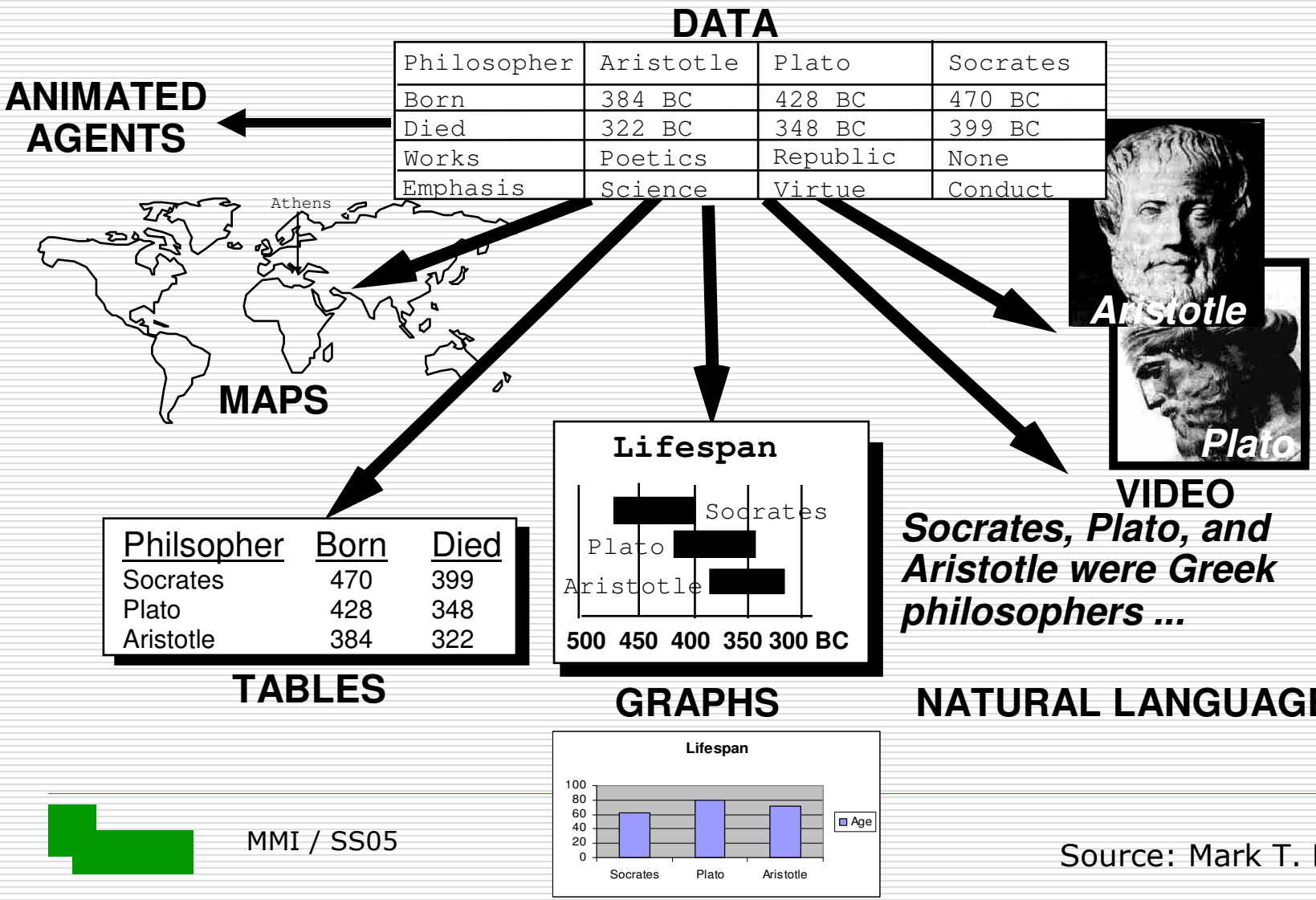
Zwei Ansätze zur Generierung multimodaler Ausgaben:

- "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

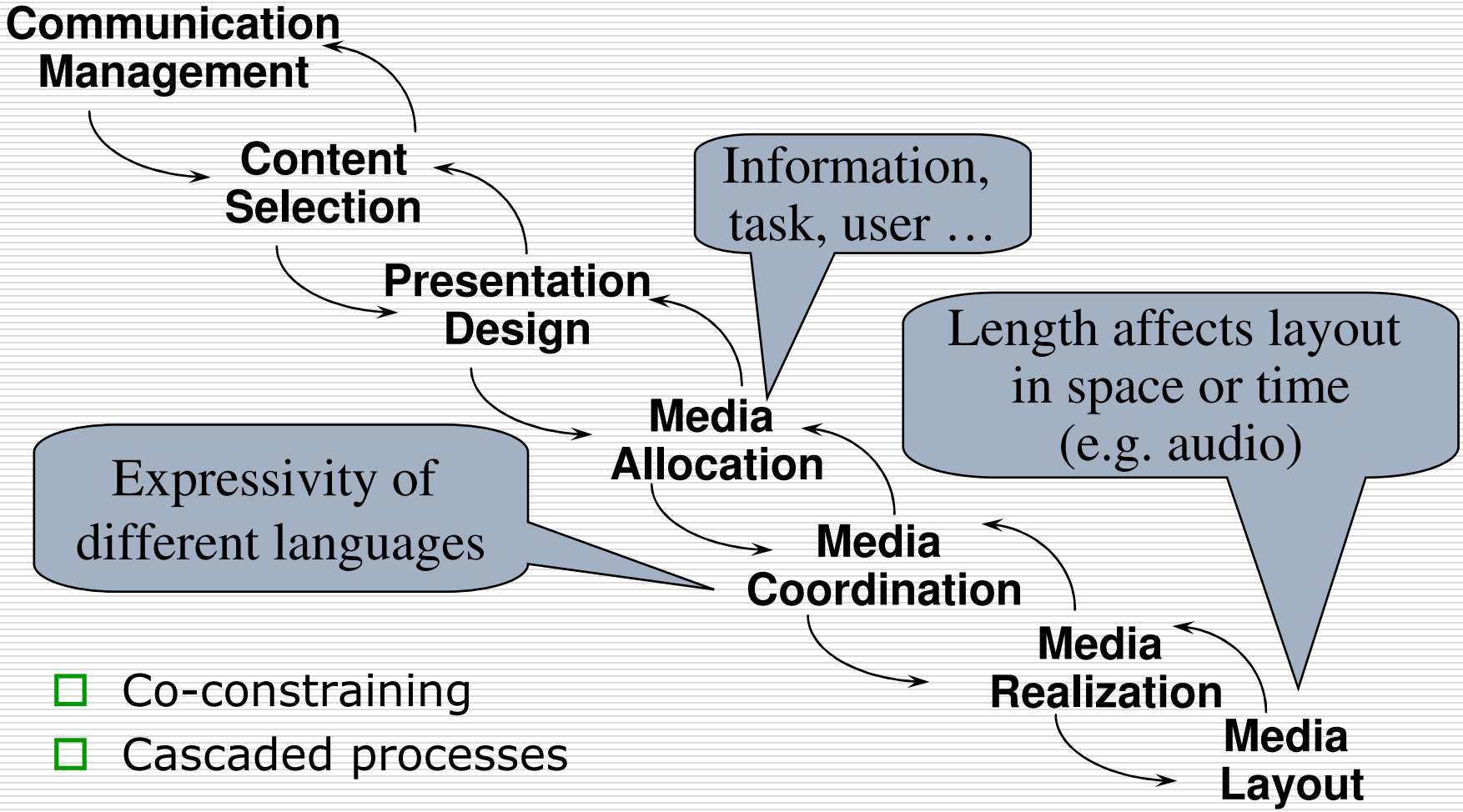
Multimedia Presentation Generation

"No Presentation without Representation"



Source: Mark T. Maybury

Common Presentation Design Tasks

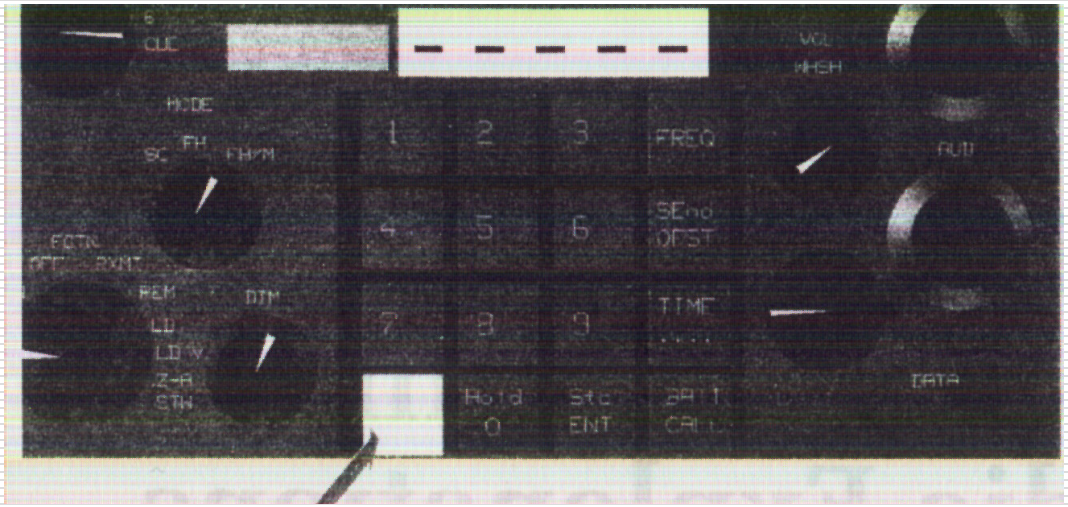


- Co-constraining
- Cascaded processes

Source: Mark T. Maybury

COMET (**Co**ordinated **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

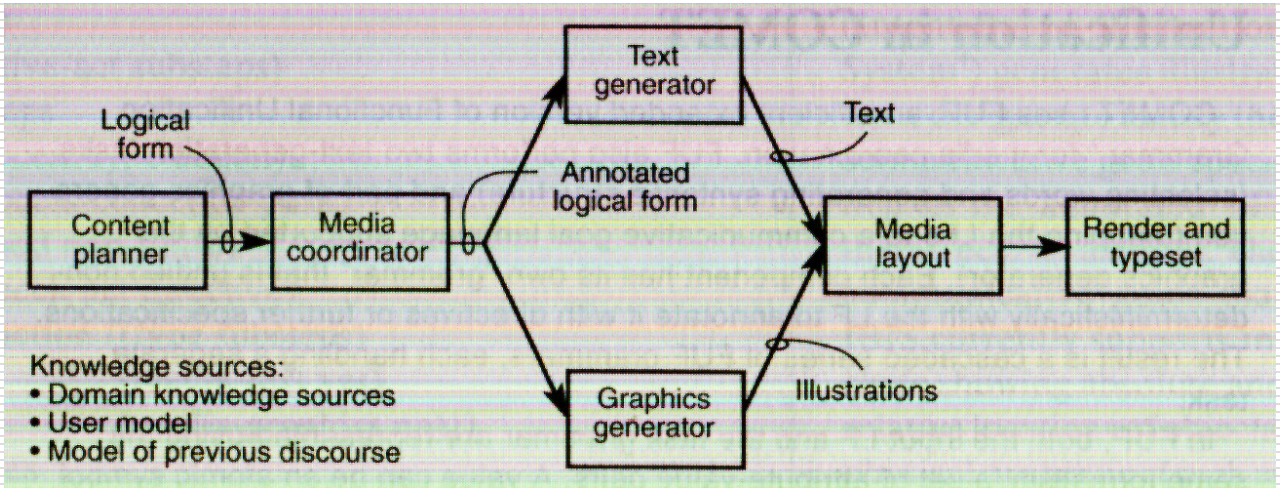


Press the CLR button to clear the display

Media coordination in COMET

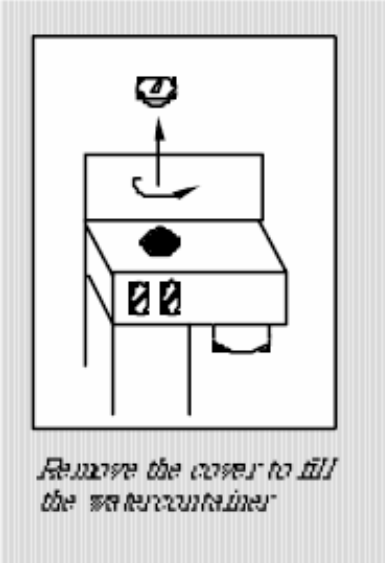
Entscheidet heuristisch, welche Information in welcher Modalität dargestellt wird

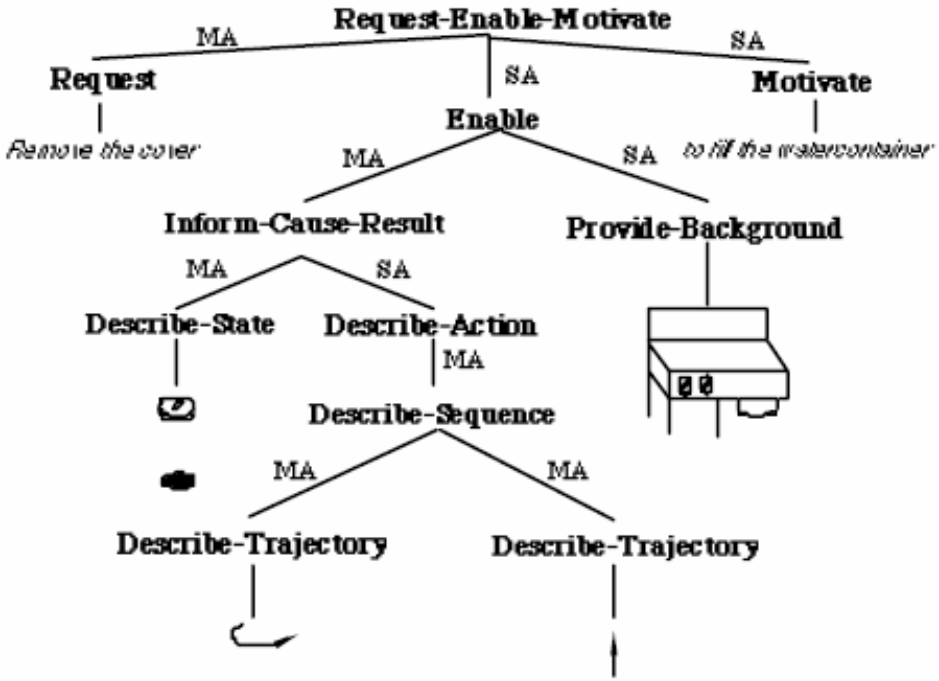
- Klassifikation der Informationen in 6 Informationstypen in der ‚logical form' für Text vs. Grafik:
  - Lokation, physische Attribute (Form etc.) → nur Grafik
  - abstrakte Aktionen und Zusammenhänge zwischen Aktionen (Reihenfolge, Kausalität) → nur Text
  - konkrete Aktionen → Grafik + Text



WIP: Use of communicative acts

- Integrated planning process to create document plan
- Use of repository 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 – anthropomorphe Agenten

- Ausgabe über "natürliche" Modalitäten
- Ein menschenähnliches, soziales(?) Gegenüber